На правах рукописи

Гомзин Андрей Геннадьевич

Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей

Специальность 05.13.11 —
«Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
кандидат физико-математических наук
Турдаков Денис Юрьевич

Москва — 2021

Оглавление

Стр.

Стр.

Стр.

Введение

В современном мире широко распространены такие способы коммуникации посредством сети Интернет, как *социальные медиа*: блоги, сайты знакомств, форумы, микроблоги, социальные сети. Особый интерес среди социальных медиа представляют социальные сети. *Социальная сеть* – платформа, онлайн-сервис и веб-сайт, предназначенные для построения, отражения и организации социальных взаимоотношений в Интернете. Основными элементами социальной сети являются публичные страницы, Они могут являться как персональными страницами пользователей, так и страницами, представляющими организации, тематические сообщества, события и т.д. Отношения между страницами представлены *социальными связями*. Примерами социальных связей являются дружба между пользователями, подписка на сообщества, события и т.д. Социальная сеть или её часть моделируется с помощью *социального графа*. Социальный граф состоит из вершин, представляющих страницы пользователей, сообществ, организаций и т.д., и рёбер, представляющих социальные связи между соответствующими вершинами.

Под *демографическими атрибутами* пользователей социальных сетей понимаются пол, возраст, семейное положение, уровень образования, род деятельности, трудоустроенность, место жительства, доход, политические, религиозные взгляды, интересы, национальность и другие. Множество значений демографических атрибутов пользователя составляют его *демографический профиль*. Множество явно указанных и публично доступных значений демографических атрибутов пользователя назовём *публичным профилем*. Не все значения указываются пользователями явно, поэтому лишь часть значений атрибутов могут быть определены с использованием публичного профиля. В связи с этим возникает задача предсказания неуказанных значений демографических атрибутов пользователей социальных медиа по доступным данным, таким как тексты публичных сообщений, социальный граф. Кроме того, некоторые пользователи преднамеренно указывают ложные данные. Отличие указанных в публичном профиле значений атрибутов от предсказанных на основе анализа поведения пользователя может служить признаком для определения ложных значений.

Для решения задачи предсказания значений демографических атрибутов необходимо специальное программное обеспечение, позволяющее собирать открытые данные из социальных сетей, применять к ним методы и модели с целью получения и восстановления демографических профилей пользователей, оценивать качество различным моделей и методов с использованием различных наборов данных. Программное обеспечение, позволяющее восстановить демографические профили пользователей, являются базовым и необходимым инструментом при решении различных прикладных задач. Так, например, значения демографических атрибутов пользователей могут использоваться коммерческими компаниями для определения целевой аудитории предлагаемых продуктов, а также для поиска потенциальных клиентов в социальных медиа. Организации могут использовать демографические профили пользователей для поиска потенциальных сотрудников с целью найма. Значения демографических атрибутов также могут быть полезными и для таких задач государственного управления, как изучение современных демографических тенденций, оценка переизбытка или нехватки специалистов в различных областях.

В диссертационная работе исследуются и разрабатываются методы и программные средства для предсказания значений *стационарных* демографические атрибутов, то есть таких, которые редко меняются и актуальны на протяжении жизни пользователей. Такими атрибутами являются пол (меняется крайне редко), год рождения (не меняется), семейное положение (в среднем меняется 1-2 раза), уровень образования (меняется по уровням, 1-2 раза), род деятельности (в среднем не меняется для взрослого человека). Методы предсказания таких атрибутов, как интересы, отношение к определённым событиям, не рассматриваются в рамках данной работы. Также в работе не рассматривается вопрос о зависимости между различными атрибутами, задача предсказания ставится независимо для каждого стационарного атрибута.

На практике одной из частых постановок задач определения значений демографических атрибутов для заданного множества целевых пользователей. Это множество может представлять собой как некоторое сообщество (студенты университета, подписчики сообщества), так и всех пользователей социальной сети. Недоступные из публичных профилей значения демографических атрибутов можно предсказывать с использованием соответствующих методов и программных средств по другим доступным данным, например, по текстам публичных сообщений пользователей. В применении к обозначенной задаче методы и про-

граммные средства для предсказания значений атрибутов по текстам имеют ряд недостатков, связанных с недоступностью, разнородностью и затруднённым сбором текстовых данных для заданного множества пользователей. Социальные связи являются более доступным источником публичных данных о пользователях. Поэтому в диссертационной работе особое внимание уделено методам, моделям и программным средствам для предсказания значений демографических атрибутов с использованием социального графа.

Существующие методы предсказания значений атрибутов по социальному графу обладают недостаточным качеством, что показывается примерами, где эти значения предсказываются неверно. В общем случае задача предсказания значений атрибутов сводится к задаче классификации или регрессии, поэтому под качеством методов понимаются традиционные для задач классификации и регрессии метрики: F-1 мера с микро- и макроусреднением, среднеквадратичная ошибка (MAE), коэффициент детерминации (R2).

Целью диссертационной работы является разработка методов и программных средств для определения значений стационарных демографических атрибутов пользователей социальных сетей. Разработанные методы должны превосходить по качеству предсказания существующие методы при доступности информации только о социальном графе.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать существующие методы определения демографических атрибутов пользователей;
2. Разработать и реализовать методы предсказания значений демографических атрибутов по социальному графу, превосходящие по качеству существующие методы;
3. Провести экспериментальное сравнение разработанных методов с существующими методами с использованием общепринятых метрик качества;
4. Разработать программную систему для определения значений стационарных демографических атрибутов пользователей социальной сети.

Научная новизна: Разработаны методы предсказания значений демографических атрибутов пользователей, основанные на введённом в диссертационной работе свойстве вершин социального графа, *специфичности контекста* для заданного атрибута. Разработанные методы показывают более высокое

качество предсказания по сравнению с методами, не использующими специфичность контекста.

Теоретическая и практическая значимость заключается в использовании разработанных методов в Talisman, комплексе взаимосвязанных программных инструментов для автоматизации типовых задач обработки данных, включая их сбор, интеграцию, анализ, хранение и визуализацию. Результаты работы были применены при выполнении работ по договору с Министерством образования и науки Российской Федерации №14.514.11.4111 «Построение социо-демографического профиля пользователей сети Интернет». Разработанные методы позволят повысить эффективность решения прикладных задач, использующих значения демографических атрибутов пользователей.

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Основные положения, выносимые на защиту:

1. Разработан подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа;
2. В рамках подхода созданы новые методы предсказания значений демографических атрибутов по социальному графу *LP-CS*, *LP-CS-Gen*, *Distr2-CS+DW[n]-XGB*, *GConv-CS[n]*, *Distr2-CS-XGB*, превосходящие по качеству существующие аналоги; даны рекомендации по их применению;
3. Реализована программная система предсказания значений атрибутов пользователей социальных сетей по социальному графу, позволившая экспериментально подтвердить превосходство созданных методов над существующими аналогами по качеству решения задачи.

Достоверность полученных результатов обеспечивается проведенной экспериментальной проверкой возможности использования специфичности контекста для предсказания значений атрибутов, с использованием данных из реальных социальных сетей, а также экспериментальным сравнением разработанных методов с аналогичными методами определения демографических атрибутов, описанными в литературе, с использованием данных из реальных социальных сетей.

Апробация работы. Основные результаты диссертационной работы докладывались в рамках следующих мероприятий:

– Научный семинар отдела «Информационных систем», Москва, 2016 г.

- 190-е заседание Московской секции ACM SIGMOD, Москва, 2016 г.
- Семинар по социофизике имени Д.С.Чернавского, Москва, 2016 г.
- Международная открытая конференция ИСП РАН 2016, Москва, 2016 г.
- 24-я международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», Москва, 2018 г.
- Ломоносовские чтения, факультет ВМК МГУ им. М.В. Ломоносова, Москва, 2020 г.
- Международная конференция «55th Annual Conference on Information Sciences and Systems (CISS)», дистанционно, 2021 г.

Публикации. Автор имеет 12 публикаций в печатных изданиях, 2 работы индексируются в Scopus и Web of science. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 3 из которых изданы в журналах, рекомендованных ВАК, 1 — в тезисах докладов. Получено 5 свидетельств о регистрации программ для ЭВМ. Основная часть работы [1] выполнена автором, редакторские правки и анализ результатов экспериментов выполнялись совместно с соавторами. Основная часть работ [2], [3] выполнена автором, редакторские правки выполнялись совместно с сооавторами. В работе [4] автором был собран набор данных, описание методов и анализ результатов был выполнен совместно с соавторами. Работа [5] полностью выполнена автором. В рамках программы [6] автором реализована часть методов сбора социальных графов. В Talisman [7] автором реализованы методы предсказания значений атрибутов пользователей социальных сетей. Большая часть программ на ЭВМ [8], [9] и [10] была реализована автором и использована для сбора данных и оценки качества работы методов.

Объем и структура работы. Диссертация состоит из введения, трёх глав, заключения и двух приложений. Полный объём диссертации составляет 143 страницы, включая 52 рисунка и 15 таблиц. Список литературы содержит 105 наименований.

Глава 1. Обзор методов определения значений демографических атрибутов пользователей

Под демографическими атрибутами понимаются пол, возраст, семейное положение, занятость, уровень образования, интересы, политические и религиозные предпочтения, национальность и другие. Пользователи в социальных сетях создают сообщения, которые могут содержать текстовую информацию, изображения, видео, гиперссылки. Исследования, описываемы в этой главе, показывают, что тематические и стилистические особенности текста сообщения определяются его автором. При этом прослеживается корреляция между характеристиками пользователя (пол, возраст, занятость) и текстами, автором которых он является. При использовании социальной сети пользователи также создают связи с другими пользователями в социальных сетях, подписываются на публичные страницы, комментируют сообщения других пользователей, отмечают сообщения, фотографии, аудио- и видеозаписи как понравившиеся. Информация о социальных связях также активно используется для предсказания характеристик пользователей.

Сначала обозреваются методы предсказания пола, возраста и других характеристик пользователей блогов и микроблогов по их текстам. Обозначаются недостатки методов в применении к задаче определения значений демографических атрибутов для заданного множества пользователей. Затем рассматриваются методы решения данной задачи без использования текстов, но с использованием информации о социальных связях. Далее описываются методы, комбинирующие структуру социального графа и текста пользователей для предсказания значений их демографических атрибутов. Кроме того, рассматриваются способы совместного предсказания значений нескольких атрибутов. Решения задач предсказания значений демографических атрибутов пользователей, описанные в главе, используют открытые данные из социальных сетей. В конце главы рассматриваются некоторые аспекты сбора таких данных.

1.1 Определение значений атрибутов пользователей по текстам их сообщений

Задача предсказания значений атрибутов по текстам сообщений в большинстве случаев сводится к задачам классификации или регрессии, в зависимости от атрибута и множества его значений. Для решения задач классификации используются методы машинного обучения с учителем. Машинное обучение с учителем позволяет найти зависимость целевых значений от исходных данных и использовать ее для предсказания значения целевого атрибута для новых данных. В нашем случае целевые данные – это значения демографических атрибутов, а исходные данные – тексты пользователей. Для использования этого подхода должна иметься выборка пользователей, для которых известны как тексты сообщений, так и значения целевых атрибутов. В процессе обучения строится модель, с помощью которой предсказываются значения атрибутов для новых исходных данных, то есть для пользователей, у которых эти значения неизвестны. При разработке методов решения прикладных задач с использованием машинного обучения с учителем выделяют несколько ключевых подзадач:

– извлечение признаков;
– отбор признаков и уменьшение размерности входных данных;
– выбор классификатора и обучение модели;

Далее описываются ранние работы, посвященные анализу текстового контента авторов, затем рассматриваются работы, посвященные предсказанию пола и возраста авторов блогов. После чего обозреваются исследования, посвящённые предсказанию демографических атрибутов пользователей микроблогов по текстам их сообщений.

1.1.1 Ранние работы

Еще до появления социальных сетей проводились исследования текстов и их авторов. При этом анализировались такие тексты, как эссе студентов, личные дневники, сообщения электронной почты (e-mail). В подобных рабо-

тах проводились статистические исследования лингвистических особенностей текстов мужчин и женщин, а также людей с различными моделями личности.

В работе [11] исследовалась зависимость между индивидуальными особенностями авторов и лингвистическим стилем авторских текстов. В качестве текстов использовались ежедневники пациентов лечебного центра, выполненные ежедневные задания студентов, аннотации научных статей известных психологов. Анализировались связь диспозициональной модели личности человека «Большая пятёрка» [12] с лингвистическими особенности текстов. Для этой цели использовалась программа LIWC (Linguistic inquiry and word count).

LIWC (Linguistic inquiry and word count) – это программа, которая осуществляет подсчет доли знаков препинания, слов с положительной и отрицательной эмоциональной окраски, слов определенных частей речи и другие признаки. Программа предназначена для анализа текстов на английском языке. Проект LIWC представляет собой один из первых шагов в изучении особенностей лингвистического стиля авторов и корреляций индивидуальных характеристик личностей и стилистическими особенностями их текстов.

Аргамон и др. [13] изучали различия между текстами, написанными женщинами и мужчинами. В работе была исследована часть корпуса BNC(British National Corpus). BNC – это корпус текстов, содержащий образцы письменного и разговорного британского английского языка из широкого круга источников. Для каждого документа известен пол автора, жанр текста, все слова размечены тэгами, обозначающими части речи. Авторы выделяли наиболее информативные части речи слов, свойственные авторам различного пола. Для определения информативности использовался алгоритм машинного обучения Balnced Winnow. Алгоритм представляет собой линейный классификатор, в котором при обучении вычисляются "веса"признаков. Данные веса показывают информативность признаков для определения, является ли автор текста мужчиной или женщиной. На этом корпусе авторы работы сделали следующие выводы: использования определений (a, the, that, these) и числительных (one, two, more, some) свойственны мужчинам; употребление местоимений (I, you, she, her, their, myself, yourself, herself) свойственно женщинам.

Ньюман и др. [14] отмечали, что в разных исследованиях анализируются тексты различных жанров и стилей. Авторы собрали вместе тексты различных жанров, различной тематики и провести анализ признаков, свойственных мужчинам и женщинам. Для обработки текстов использовалась программа LIWC,

описанная выше. Авторы пришли к следующим выводам: женщины использовали больше слов, связанных с психологическими и социальными процессами. Мужчины больше ссылались на свойства объектов и безличные темы.

1.1.2 Анализ текстов блогов и электронной почты

В начале 2000х годов начали набирать популярность блоги. Блог (англ. blog, от web log – интернет-журнал событий, интернет-дневник, онлайн-дневник) – веб-сайт, основное содержимое которого – регулярно добавляемые записи, содержащие текст, изображения или мультимедиа. В это время задача формулировалась как предсказание неизвестных значений демографических атрибутов пользователя блога по текстам его авторства. При этом в качестве текстов рассматривались тексты сообщений электронной почты и посты в блогах. Чаще всего встречаются работы, посвященные таким атрибутам, как пол и возраст. В эти годы стали появляться работы, решающие задачу предсказания атрибутов с помощью машинного обучения с учителем.

Де Вел и. др. [15] исследовали связь языковых особенностей сообщений электронной почты (e-mail) со значением пола их авторов. Рассматривались тексты писем на английском языке. Пол авторов сообщений предсказывалсяы с использованием метода опорных векторов (SVM, англ. Support Vector Machine). На вход классификатору SVM подаётся комбинация стилометрических, структурных и гендерно-специфичных признаков. Примерами стилометрических и структурных признаков являются количество пустых строк в тексте письма, частота служебных слов, количество пробелов, табуляций, заглавных символов, количество приложенных к письму файлов. В качестве гендерно-специфичных признаков авторы использовали количество слов, оканчивающихся на «-able», «-al», «-ful», и др., количество слов «sorry» и слов, начинающихся на «apolog-». Для оценки данного метода использовался набор данных, состоящий из 4369 сообщений от 325 различных авторов. Было достигнуто более 70% F1-меры.

Херринг и др. [16] проанализировали данные из 100 блогов, собранных с сайта blo.gs в 2004 году. Исследовались зависимости между языковыми особенностями текстов постов и такими характеристиками, как пол автора и жанр поста. Под жанром понимается тексты о жизни самих авторов и тексты и

внешних к авторам событиями. Авторы выбрали потенциальные признаки, свойственные для каждого пола и жанра. Среди них такие английские слова, как I, me, my, mine, we our, ours, let's, she, he, they, them, their, theirs, числительные и др. Для статистического анализа использовалась модель логистической регрессии. В результате анализа авторы сформулировали вывод о том, что предложенные признаки незначимы для разделения авторов по полу, однако хорошо разделяют тексты по рассматриваемым жанрам.

В работе [17] Бургер и Хендерсон решали задачу предсказания возраста пользователей блогов. Задача ставилась как бинарная классификация: моложе 18 лет, 18 лет и старше. Авторы проанализировали 100000 постов блогов и изучили признаки, которые потенциально можно использовать для предсказания возраста. В качестве признаков, извлекаемых из текстов, использовались длина сообщения, доля знаков препинания в текстах пользователя, последовательности слов и символов (n-граммы), количество гиперссылок в тексте. Помимо текстовых признаков использовалось время суток публикации сообщения, а также информация из профиля: страна, количество друзей, интересы. Использование регрессионной модели незначительно повысило качество работы простого базового решения, всегда возвращающего самый частый класс.

Счлер и др. [18] предсказывали значения пола и возраста с использованием машинного обучения с учителем. Значения возраста разбивались на интервалы: 13-17, 18-22, 23-27, 28-32, 33-37, 38-42, 43-48, старше 48. В качестве признаков выбраны части речи, служебные слова, гиперссылки, 1000 наиболее информативных в тренировочном наборе юниграмм. В качестве значений признаков использовалась частота признака в соответствующем тексте. В качестве алгоритма машинного обучения для классификации использовался алгоритм Multi-Class Real Winnow (MCRW) [19] На наборе данных, включающем в себя тексты из 71000 блогов авторами достигнуто качество предсказания 70-80% для обоих рассматриваемых атрибутов.

Ян [20] использовал наивный байесовский классификатор для предсказания пола авторов постов блогов. Авторы предлагают помимо стандартных юниграмм использовать такие признаки, как цвет фона, шрифты, знаки препинания, эмотиконы. Эксперименты проводились на данных, полученных из блогов Xanga. Набор данных содержит 75000 постов от 3000 блоггеров. По результатам экспериментов были сделаны следующие выводы: дополнительные признаки позволяют улучшить качество предсказания пола пользователей;

удаление признаков, соответствующих стоп-словам, лишь ухудшает качество предсказания.

Новсон и Оберландер [21] определяли пол автора блога по текстовому содержимому. Использовалось три типа признаков: полученные с использованием программы LIWC, полученные помощью базы данных MRC [22] и n-грамм. В качестве алгоритма машинного обучения использовался классификатор SVM. Результаты экспериментов показали, что качество предсказания при использовании признаков, учитывающих контекст слов, т.е. n-грамм, существенно выше, чем при использовании признаков, не использующих контекст слов. Также авторы провели эксперимент с уменьшением количества n-грамм: были оставлены лишь наиболее значимые для пола автора признаки. Было показано, что качество предсказания в этом случае не уменьшается, однако вычисления существенно ускоряются.

Ченг и др. [23] рассматривают задачу предсказания пола автора по относительно короткому тексту. Авторы проводят эксперименты на двух наборах данных. Первый содержит короткие новостные статьи (около 500 слов), написанные журналистами одного из онлайн-изданий. Второй набор данных представляет собой набор e-mail сообщений, средняя длина которых около 115 слов. Для предсказания пола автора сообщения используется три модели машинного обучения: байесовская логистическая регрессия, ансамбль деревьев принятия решений и классификатор SVM. В качестве признаков рассматриваются признаки на уровне символов (общее количество символов, количество букв, цифр, пробельных и специальных символов), признаки на уровне слов (общее количество слов, средняя длина слова, признаки LIWC и др.), синтаксические признаки (количество кавычек, вопросительных знаков, и др.), структурные признаки (количество строк, предложений, среднее количество слов в абзаце и др.), служебные слова (количество артиклей). В результате использования классификатора SVM получены максимальные значения точности предсказания пола.

Нгуыен и др. [24] решали задачу предсказания возраста по текстам. Рассматривались три корпуса текстов: тексты блогов, корпус с транскрипцией телефонных разговоров, посты пользователей с форума, посвященного раку груди. Метод предсказания возраста заключается в применении алгоритма линейной регресии. В качестве признаков для линейной регрессии авторы использовали юниграммы слов, юниграммы и биграммы частей речи для слов,

полученные с помощью программы для извлечения частей речи Stanford POS tagger [25], классы слов, полученные с помощью LIWC. Значением признака являлась частота признака, нормализованная в рамках одного документа (текста).

1.1.3 Анализ текстов пользователей микроблогов и социальных сетей

В конце 2000х годов возросло количество пользователей микроблогов. Тексты сообщений, которые пользователи публикуют в микроблогах, обладают некоторыми существенными для анализа особенностями. Одной из главных особенностей является короткая длина сообщения. Как правило, одно сообщение несет в себе одну законченную мысль и состоит из одного-двух предложений. Кроме того, сервисы микроблогов зачастую ограничивают максимальную длину сообщений (например, в Twitter максимально возможная длина сообщения равна 140 символам). Сообщения часто содержат орфографические ошибки.

В отличие от текстов блогов, которые готовятся, проверяются и вычитываются перед публикацией, целью сообщений в микроблоге является максимально быстро и коротко выразить некоторую короткую мысль. Многие пользователи не проверяют свои сообщения на наличие орфографических ошибок, отправляют их сразу после набора. Кроме того, некоторые ошибки делаются преднамеренно: если длина сообщения больше допустимой, пользователь сокращает слова.

Сообщения в микроблогах представляют собой новости, мнения, которые имеют некоторую эмоциональную окраску. Пользователи передают свои эмоции с помощью эмотиконов (сокращения и значки для обозначения эмоций, напрмер, «:-)»), повторяющихся символов и знаков препинания (например, «AAAA», «!!!!»).

С появлением микроблогов появились и свои стандарты и специальные символы. Например, символ @, после которого следует имя пользователя микроблога. Также, в микроблоге Twitter появился специальный символ, «#», который означает хэштеги – специальные ключевые слова, указываемые пользователями.

Далее описываются методы, посвященные предсказанию демографических характеристик по коротким текстам с перечисленными выше особенностями. Зачастую в целях улучшения качества работы на подобных текстах, используются дополнительные доступные данные, например, из профилей пользователей.

Бургер и др. [26] предсказывали пол пользователей социальной сети и сервисе микроблоггинга Twitter по текстам их сообщений. Сообщения в этой социальной сети называются твитами. Для предсказания пола пользователей использовались символьные и словесные n-граммы из твитов, поля «о себе» профиля, полного имени и короткого имени (никнейма). Тексты разбивались на слова с помощью простого метода токенизации, разделяющего слова в тех местах, где происходит смена алфавитного и неалфавитного символов. В качестве классификаторов использовались наивный байесовский классификатор, SVM с линейным ядром и Balanced Winnow 2 [27].

Цоновер и др. [28] определяли политические предпочтения пользователей социальной сети Twitter. Рассматривались три класса: демократы, республиканцы, неявная политическая позиция. В качестве признаков, извлечённых из текстов сообщений, использовались юниграммы слов и хэштеги. Дополнительно использовались признаки, извлёченные из структуры сети ретвитов сообщений. Эти признаки представляют собой кластеры, извлечённые из структуры графа с помощью метода на основе распространения меток [29]. В качестве модели для классификации авторы выбрали SVM. Для оценки качества предложенного метода авторы собрали набор данных. Набор данных собирался с помощью ручной разметки 1000 пользователей, активно участвующих в обсуждении политики США. На полученном наборе данных авторы достигли следующих результатов: при использовании только хэштэгов точность достигла 90.8%, при использовании всех признаков – 94.9%.

Рао и др. [30] рассматривали задачи предсказания значений различных атрибутов пользователей Twitter: пола, возрастного интервала (< 30 лет, $\geqslant 30$ лет), политических взглядов (республиканцы, консерваторы), региона (южная и северная Индия). Предсказание значений каждого из атрибутов рассматривалась как независимая задача. Референсные значения для пользователей были извлекались вручную. Для каждого значения атрибута было собрано от 200 до 1000 пользователей. Рассматривалось два вида признаков, извлекаемые из текстов твитов. Социо-лингвистические признаки включают в себя зара-

нее заданные эмотиконы, повторяющиеся знаки препинания для выражения эмоций (например, «!!!!!!!», «!?!!??!»), повторяющиеся символы в словах (например, «noooooo») и т.д. В качестве второй группы признаков рассматривались юниграммы и биграммы слов, взвешенные нормализованной частотой. В качестве классификатора авторы использовали SVM. Классификатор применялся для каждой из группы признаков. Для использования обеих групп признаков применялся один из методов ансамблирования – стекинг. На вход еще одному классификатору SVM подавались выходы классификаторов по социо-лингвичтическим признакам и по n-граммам. Результаты экспериментов показали, что такой подход показывает большее качество, чем отдельные классификаторы по каждой из группе признаков.

Пирсман и др. [31] рассматривали задачу определения пола и возраста авторов коротких сообщений из социальной сети Netlog. Рассматривались две категории возраста: $\leqslant$ 16 лет и $\geqslant$ 25 лет. Авторы применяли метод SVM для классификации сообщений. Для обучения и оценки качества был собран корпус сообщений на фламандском голландском языке. Предварительно тексты были разбиты на токены и предобработаны. Отдельно выделялись токены для эмотиконов и пунктуации. Кроме того, авторы использовали нормализацию слов, что позволило исправить некоторые орфографические ошибки. В качестве признаков использовались n-граммы (n от 1 до 3) на уровне символов и токенов. Особое внимание в данной работе уделяется отбору наиболее информативных признаков для предотвращения эффекта переобучения. В качестве метода отбора признаков авторы используют критерий хи-квадрат [32].

Из-за увеличения словаря и уменьшения размера текстов, методы на основе машинного обучения сталкивались с проблемой переобучения. Для преодоления переобучения применялась предобработка признаков. При этом использовались как методы отбора наиболее информативных признаков, так и способы проецирования исходных признаков на пространство более низкой размерности. Во втором случае разреженные представления в пространстве высокой размерности проецировались на плотные представления в новом пространстве более низкой размерности.

Работы Деитрик, Миллер и др. [33; 34] посвящены предсказанию пола авторов твитов. В работах авторы проводят отбор признаков, включающих в себя n-граммы символов и слов. Для отбора используются несколько методов, включая Хи-квадрат, Information Gain, Information gain Rate, Relief, Symmetrical

Uncertainty, Filtered Attribute Evaluation. Для получения окончательного результата отбора признаков использовалось голосование. В качестве алгоритмов для классификации с использованием отобранных признаков использовались алгоритмы Перцептрон, Наивный байесовский классификатор, модифицированная нейронная сеть Balanced Winnow.

Работа Преотиуц-Пиетро и др. [35] посвящена предсказанию рода деятельности пользователей Twitter. В качестве возможных значений рода деятельности рассматриваются значения из SOC[1]. Референсные значений рода деятельности извлекаются из профилей пользователей эвристическим методом. В поле «о себе» пользователей ищутся заранее заданные шаблоны текста, указывающего на тот или иной род деятельности. Затем собранные профили были вручную просмотрены и отфильтрованы неверно извлечённые значения. В отличие от работ, описанных выше, вместо отбора признаков авторы рассматривают методы уменьшения размерности данных. В данном случае исходное (как правило, разреженное) представление объекта выборки, трансформируется и представляется в новое (как правило, плотное) представление в новом признаковом пространстве. Рассматриваются несколько способов представления слов. Первое представление получается с помощью применения сигулярного разложения матрицы, элементами которой являются значения поточечной взаимной информации [36]. Друге представление строится с использованием спектральная кластеризации [37] той же матрицы. Авторы также рассматривают векторные представления (так называемые «вложения»), полученные методом Word2vec, предложенным Микиловым [38] и кластеры, полученные из этих представлений. Признаковое представление пользователей получается суммированием представлений используемых им слов. В качестве классификаторов авторы рассматривали Гауссовский процесс, линейную регрессию, SVM с ядром RBF. Наилучшее качество было достигнуто при использовании Гауссовского процесса с кластерами, полученными с использованием представлений слов word2vec.

Пандя и др. [39] рассматривали задачу предсказания возраста пользователей Twitter. Особое внимание авторы уделяли хэштэгам и гиперссылкам в текстах сообщений: для них извлекается дополнительная информация, контекст. Контекст хэштэга формировался из текстов твитов, в которых встречается заданный хэштэг. Для гиперрссыллки контекстом являлся заголовок

[1]Standard Occupation Classification

страницы, на которую она ссылается. Текст сообщения, а также контекст, извлечённый из хэштэгов и гиперссылок далее разбивался на слова, каждому слову ставился в соответствие вектор, полученный заранее обученной моделью Word2vec [38]. К полученному представлению применяется свёрточная нейронная сеть. Авторы экспериментально показали, что описанный метод показывает более высокое качество по сравнению с методом, основанном на использовании классификатора SVM над n-граммами.

Одним из подходов, позволяющим уменьшить размерность признаков, полученных из текстов, является тематическое моделирование [40—42]. Этот подход при определении тем для документа учитывает синонимию и омонимию, то есть одинаковые значения для различных слов и различные значения для одного слова. Тематическое моделирование находит применение в различных прикладных задачах обработки текстов.

Утеуов [43] применял подход на основе вероятностного тематического моделирования для предсказания интересов пользователей социальных сетей. По текстам пользователей и других публичных страниц с помощью модели ARTM, разработанной под руководством Воронцова [42], строится тематическая модель, которая затем используется для предсказания интересов. К качестве значений интересов рассматриваются значения, автоматически извлечённые из профилей. В работе также применятся подход, в котором используются извлечённые темы для групп, на которые подписан пользователь, что позволяет применять метод даже при небольшом количестве текстов пользователей.

Смелик и Фильченков [44] применяли тематическое моделирование для задач автоматического аннотирования изображений и иллюстрирования текстов. Предложенный метод может быть использован для получения текстовых представлений публикуемых пользователями изображений. Это позволит применять методы предсказания значения атрибутов по текстам, даже если пользователь предпочитает публиковать изображения вместо текстов.

В социальных сетях публичные страницы часто делятся на несколько типов. Помимо публичных страниц обычных пользователей выделяются специальные тематические страницы, которые объединяют пользователей по интересам, формируя сообщества пользователей. Такие страницы могут представлять организации, официальных лиц, знаменитостей. Встречаются также страницы, посвященные определённым мероприятиям. В некоторых социальных сетях (в частности, Twitter) аккаунты не разделяются по типам. При

предсказании значений демографических характеристик важно понимать типы аккаунтов, так как данные характеристики имеют смысл только для аккаунтов, представляющих пользователей. Тип страницы также можно рассматривать как атрибут, который можно предсказать.

Любевсиц и Фисер [45] рассматривали задачу определения типа аккаунта. Под типом понимается, является ли аккаунт частным или корпоративным. Для решения данной задачи используется классификатор SVM с ядром RBF. В качестве признаков рассматривались как независимые от языка сообщений признаки так и языкозависимые. В качестве первых использовались количество твитов, содержащих гиперссылки, средняя длина твита, количество ретвитов, и т.д. В качестве языкозависимых признаков рассматривались части речи слов, а также наиболее информативные слова. Для оценки качества предлагаемого метода авторами был собран набор данных из 7,5 миллионов сообщений на словенском языке от 7778 авторов (аккаунтов). Тип аккаунтов размечался вручную.

В рамках диссертационной работы был выполнен обзор методов определения демографических атрибутов пользователей по текстам их сообщений, который опубликован в работе [2].

1.1.4 Экспериментальное сравнение методов предсказания значений атрибутов пользователей по текстам комментариев в социальной сети

В рамках диссертационной работы было проведено экспериментальное сравнение различных методов предсказания значений возраста и уровня образования по текстам публичных комментариев пользователей. Методы были выбраны и реализованы студентами МГУ и ВШЭ в рамках практической части курса по обработке текстов, проводимого Турдаковым Д.Ю.[2] Цель данного курса – дать студентам необходимые теоретические значения для решения открытых проблем обработки естественного языка. В рамках практической части курса предлагалось решить одну из прикладных задач обработки текстов. В качестве такой задачи осенью 2017 года рассматривалась задача предсказания

[2] http://tpc.at.ispras.ru

уровня образования и возраста пользователей социальной сети Вконтакте по текстам их комментариев. Для оценки полученных решений был собран набор данных, часть которого была доступна студентам. Студенты предлагали свои решения, которые затем оценивались с использованием этого набора данных. Сначала описывается набор данных, затем постановка задачи, предложенные студентами решения и результаты экспериментальной оценки качества решений.

Описание набора данных

Набор данных был собран из комментариев пользователей к сообщениям, публикуемых на публичных страницах тематических сообществ социальной сети Вконтакте. Для сбора данных было выбрано 1000000 наиболее активных публичных страниц. Были собраны тексты комментариев, которые затем были сгруппированы по авторам.

Каждый пользователь в наборе данных представлен анонимизированным профилем и множеством текстов комментариев. В набор данных вошли только пользователи с не менее 20 собранных комментариев на русском языке.

Профиль пользователя состоит из значений двух атрибутов: возраст и уровень образования. Значения возраста были разделены на следующие интервалы: «$\leqslant$ 17», «18 – 24», «25 – 34», «35 – 44», «$\geqslant$ 45». Задача заключалась в предсказании возрастного интервала пользователей. Уровень образования представлен тремы возможными значениями: «ниже среднего», «среднее», «высшее». Под пользователями с «высшим» образованием подразумеваются те, кто окончили вуз; пользователи с уровнем образования «ниже среднего» на данный момент учатся в средних школах; уровень образования «среднее» означает, что пользователь окончил среднюю школу, но не окончил вуз, в частности, в эту же категорию попадают студенты вуза. Эти значения извлекались из открытых значений, указанных на персональных страницах пользователей. В набор данных включены пользователи, у которых указано хотя одно из значений: возраст или уровень образования. Пользователи с возрастом больше 18 лет и уровнем образования «ниже среднего» не были включены в набор данных. Аналогично, в набор данных не были включены пользователи до 15 лет, у которых

уровень образования указан как «высшее» или «среднее». Такие комбинации предполагаются маловероятными, наличие их скорей всего связано с неверно указанными в профиле данными.

После сбора набора данных, он был разбит на три непересекающиеся части: тренировочная, тестовая-1, тестовая-2. Тренировочная часть состоит из 8607 пользователей, тестовая-1 состоит из 1070 пользователей, тестовая-2 состоит из 1053 пользователей. Тренировочная часть была дана студентам и использовалась для обучения предлагаемых моделей. Тестовая-1 и тестовая-2 использовались для оценки качества решений. Тестовая-1 использовалась для еженедельной оценки, тестовая-2 – для финальной.

Постановка задачи и ограничения

Участникам практической части курса было предложено реализовать решения двух задач:

- Предсказание возрастного интервала пользователя по его текстам. На вход подаётся список текстов одного автора. На выходе ожидается один из следующих классов: «$\leqslant 17$», «$18-24$», «$25-34$», «$35-44$», «$\geqslant 45$»;
- Предсказание уровня образования пользователя по его текстам. Вход аналогичен предыдущей задаче. На выходе ожидается один из следующих классов: «ниже среднего», «среднее», «высшее».

Авторами курса было подготовлено два базовых решения. Первый всегда возвращает фиксированное значение, являющееся самым частым в наборе данных: «25-34» для возраста, «высшее» для уровня образования. Второй базовый метод представляет линейный SVM, который в качестве признакох получает слова (юниграммы), взвешенные методом TF-IDF. Оказалось, что простое второе базовое решение превосходит другие классификаторы с этими признаками, а также многие решения с другими признаками. Участники получали 1 балл, если их решение превосходит по качеству первое базовое решение и 2 балла, если решение превосходит оба базовых решения. Для мотивации студентов на улучшение своих решений, даже если результат оказался лучше второго базового решения, были введены дополнительные баллы. Решения студентов были отранжированы по качеству на тестовом наборе. Первые 8 студентов получали

дополнительные баллы в соответствие с рангом, если их решение превосходит оба базовых решения. Первый в рейтинге получает 8 дополнительных баллов, второй – 7, и т.д. Каждый студент видел свою позицию в рейтинге. Полученные финальные баллы учитывались при выставлении оценки за курс.

В отличие от других соревнований, проводимых, например, Kaggle[3], решения запускались на специально выделенной машине с ограниченными физическими ресурсами и программным окружением. Это позволяет дать равные условия участникам, а также даёт возможность не открывать тестовые наборы для студентов. Автоматическая система тестирования имела следующие ограничения. Поддерживался язык программирования Python 3.x. Допускалось использование общеиспользуем библиотек для языка Python (NLTK [46], scikit-learn [47], pythorch [48], keras [49] и др.). Архив с решением не должен превышать 15 Mb, время работы решения в рамках тестирования ограничено 20 минутами, используемая оперативная память ограничена 16 Gb. Участникам разрешалось присылать предобученные модели. Каждый участник мог прислать не более 10 решений в неделю. Общая продолжительность сдачи практического задания – 9 недель. Из-за ограничений количества посылок, участникам рекомендовалось использовать метод скользящего контроля на доступной им тренировочной части данных для поиска наиболее производительных гиперпараметров перед отправкой решения.

Дополнительно была обучены модели fasttext [50] с размерами векторов 100, 200 и 300. Модели обучалась на 3.3GB данных, содержащих сообщения и комментарии пользователей Вконтакте. Этими моделями можно было пользоваться в решениях.[4]

Предложенные участниками решения

Всего было прислано 209 версий решения для задачи предсказания возраста и 179 решений для задачи предсказания уровня образования.

В представленных решениях использовались линейный SVM, логистическая регрессия, наивный байесовский классификатор, пассивно-агрессивный

[3] https://www.kaggle.com/

[4] http://tpc.at.ispras.ru/dopolnitelnye-materialy/

Таблица 1 — Список лучших 10 методов предсказания возраста и уровня образования по текстам комментариев пользователей

№	Классиф-р	n-граммы	n	Токенизация	Особенности
1	Лин. SVM	char_wb	1-4	по умолчанию	C=2.5
2	Лин. SVM	word	1	NLTK	sublinear_tf max_df=0.95 C=1 Стемминг
3	Лин. SVM	word	1	NLTK	sublinear_tf C=1
4	LSTM	word	1	по умолчанию	max_feat=10^5
5	Лин. SVM	char	2-4	по умолчанию	ё → e sublinear_tf C=1
6	Лин. SVM	word	1	TweetTokenizer	sublinear_tf C=1
7	Голосование	char	3-4	TweetTokenizer	Стоп слова LOG C=148 SVM C=0.68
8	Лин. SVM	word	1-4	TweetTokenizer	max_feat=$5 * 10^5$ C=15
9	Голосование	char	3	TweetTokenizer	LOG C=75, SVM C=0.8
		word	1		
10	Лог. регр.	char	3	по умолчанию	C=17 word max_feat=$2 * 10^5$ char max_feat=$2 * 10^5$ Стемминг
		word	1		

классификатор, стохастический градиентный спуск, метод регуляризации Тихонова и нейронные сети. В качестве признаков использовались словесные и символьные n-граммы, взвешанные с помощью TF-IDF или ненормализованной частоты. Также использовались вектора, полученные методом fasttext.

Детали 10 наилучших решений представлены в таблице 1. лучшие решения основаны на *Линейном SVM*, *LSTM* [51] и *голосовании*. Классификатор, основанный на голосовании, использовал логистическую регрессию и линейный

SVM в качестве моделей, а затем выбирал окончательный ответ в соответствие с максимальной вероятностью. C – гиперпараметр логистической регрессии и линейного SVM.

Столбцы *n-граммы* и *n* описывают тип и длину n-грамм, соответственно. Тип *word* означает n-граммы слов, тип *char* означает символьные n-граммы с учетом пробельных символов, тип *char_wb* означает символьные n-граммы, извлеченные из отдельных слов. Параметр *sublinear_tf* обозначает решения, в которых частота n-граммы tf заменяется на $1 + \log(tf)$ при вычислении весов TF-IDF. Параметр *max_df* означает порог, позволяющий игнорировать токены, частота которых в документе строго превышает его. *ё* → *е* означает предобработку, в которой все символы «ё» заменяются на «е». Токенайзер по умолчанию означает стандартную токенизацию, используемую в библиотеке scikit-learn. В двух решениях используется *стемминг* методом Портера [52]. *TweetTokenizer* – специальный токенайзер, созданный для извлечения токенов из сообщений социальной сети Twitter, реализованный в библиотеке NLTK. Токенайзер *NLTK* означает токенайзер, используемый по умолчанию в библиотеке NLTK.

Все представленные в таблице 1 решения использовали n-граммы со взвешиванием TF-IDF в качестве признаков.

Не смотря на ожидания, решения, использующие вектора fasttext не показал приемлемого качества результаты и не были включены в таблицу 1. Также студенты использовали методы отбора наиболее репрезентативных признаков, однако такие решения не показали высокого качества предсказания значений возрастного интервала и уровня образования пользователей.

Результаты

Для оценки качества использовалось значение точности (accuracy), оцененное на двух тестовых частях набора данных:

$$accuracy(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \tag{1.1}$$

где $\hat{y}_i$ – предсказанное значение для пользователя i, y_i – истинное значение, $n_{samples}$ – количество пользователей, для которых измерялось качество.

Значения точности методов, представленных в таблице 1, на наборах тестовый-1 и тестовый-2 приведены в таблице 2. В таблицу также включены базовые решения и соответствующие значения точности. Решения отсортированы по убыванию точности отдельно для каждой задачи (возраст и образование) и тестового набора.

Таблица 2 — Качество методов предсказания возраста и образования по текстам комментариев пользователей (отсортированы для каждого из тестовых наборов). bl_1 и bl_2 – базовые решения

Возрастной интервал				Уровень образования			
тестовый-1		тестовый-2		тестовый-1		тестовый-2	
Реш.	Точ.	Реш.	Точ.	Реш.	Точ.	Реш.	Точ.
1	0.61460	8	0.59656	5	0.7712	9	0.76671
2	0.61204	3	0.58995	1	0.76485	7	0.76419
3	0.61075	5	0.58995	6	0.76485	5	0.76167
8	0.60819	2	0.58862	7	0.75979	10	0.76167
4	0.60179	1	0.56614	9	0.75095	6	0.75284
5	0.60179	4	0.55423	10	0.74716	1	0.7465
bl_2	0.58259	bl_2	0.56481	bl_2	0.73957	bl_2	0.73140
bl_1	0.31626	bl_1	0.30556	bl_1	0.50569	bl_1	0.47793

Наилучшие результаты были достигнуты методами, основанными на линейном классификаторе SVM. В 10 лучших решений попало также решение, основанное на модели LSTM [51]. Стоит отметить, что LSTM показывает хорошие результаты только на наборе «тестовый-1» для задачи предсказания возраста. На наборе «тестовый-2» точность оказалась ниже второго базового решения. Прочие решения, использующие нейронные сети показали неудовлетворительные результаты.

Три из десяти методов (2, 3, 8) входят в 4 лучших решений для обоих тестовых наборов для задачи предсказания возраста; 2 из 10 метода (5, 7) входят в 4 лучших решения для обоих тестовых наборов для задачи предсказания уровня образования. 5 из 10 лучших решений использовали символьные n-граммы, два из них входят в 4 лучших метода одновременно для двух задач. Первые 3 лучших решения использовали линейный SVM для обоих задач.

Выводы и анализ результатов

Было проведено экспериментальное сравнение методов машинного обучения для задач предсказания возраста и уровня образования пользователей социальных сетей по текстам их комментариев. Поиск лучших методов проводился среди решений, предложенных группой студентов, являющихся частниками курса по обработке текстов на естественном языке. Одним из лучших подходов оказалось использование в качестве классификатора линейного SVM. В качестве признаков использовались словесные и символьные n-граммы, взвешенные с помощью TF-IDF.

Большинство современных подходов, основанных на нейронных сетях, не показали удовлетворительных результатов. Есть две гипотезы, с чем это может быть связано. Первая заключается в ограничении на вычислительные ресурсы, доступные для запуска решений. Вторая возможная причина – нестабильность алгоритмов на зашумлённых данных. В качестве референсных значений атрибутов использовались значения из профилей, указанные самими пользователями. Пользователи могут указывать ложную информацию в профиле. При подготовке набора данных была выполнена лишь простая фильтрация по соответствию значений возраста и уровня образования.

С учетом условий и ограничений практической части курса, линейный SVM показал себя как хороший подход для решения поставленных задач.

После окончания практической части курса, полученные результаты были проанализированы и опубликованы в работе [4].

Стоит отдельно отметить, что в рамках данной работы сбор данных проводился не таргетированно, то есть данные собирались не для заданных пользователей, а доступные случайные сообщения. В набор данных вошли лишь пользователи, которые оставили достаточное количество публичных сообщений на просмотренных страницах.

По результатам исследования можно выделить ряд недостатков использования текстов комментариев для предсказания значений демографических атрибутов для заданного множества целевых пользователей. Таргетированный сбор текстов комментариев для заданного множества пользователей социальной сети часто затруднён, так как комментарии могут располагаться на различных публичных страницах и относиться к различным публикациям. Не все соци-

альные сети предоставляют возможность получить множество всех сообщений для заданного пользователя. Например, популярная социальная сеть Вконтакте не даёт такой возможности. Кроме того, не все пользователи активно пишут публичные комментарии. Как следствие, для многих пользователей не удаётся собрать достаточное количество текстов, что оказывает существенное влияние на качество методов предсказания значений атрибутов по текстам комментариев. Предсказание значений атрибутов с использованием социальных связей лишено обозначенных недостатков.

1.2 Определение значений атрибутов пользователей по социальным связям

В этом разделе обозреваются работы, посвященные задаче предсказания значений демографических атрибутов пользователей по заданному социальному графу. Задача ставится следующим образом. Рассматривается один атрибут. По заданной части социальной сети в виде социального графа и известным значениям атрибута для некоторого подмножества пользователей предсказать значения атрибута для остальных пользователей.

Для некоторых атрибутов применимы методы, оценивающие значение атрибута по значениям соседей в социальном графе, например, среднее или наиболее частое значение. Такие методы работают в случае наличия «гомофилии» для предсказываемого атрибута в рассматриваемой части социальной сети и достаточного для этого количества данных. Авторские исследования [3], проводимые в рамках диссертационной работы, показали применимость такого подхода для предсказания возраста пользователей социальной сети Вконтакте. На рисунке 1.1 представлено эмпирическое распределение значений возраста для пользователей, связанных отношением дружбы. Более тёмной и синей ячейке соответствует бóльшая вероятность. Каждая строка представляет собой распределение возраста соседей для пользователей с фиксированным возрастом. Стоит отметить, что распределение несимметричное, так как при построении графика использовались только доступные данные о соседях. Так, например, для двух пользователей, являющихся друзьями, у одного из которых

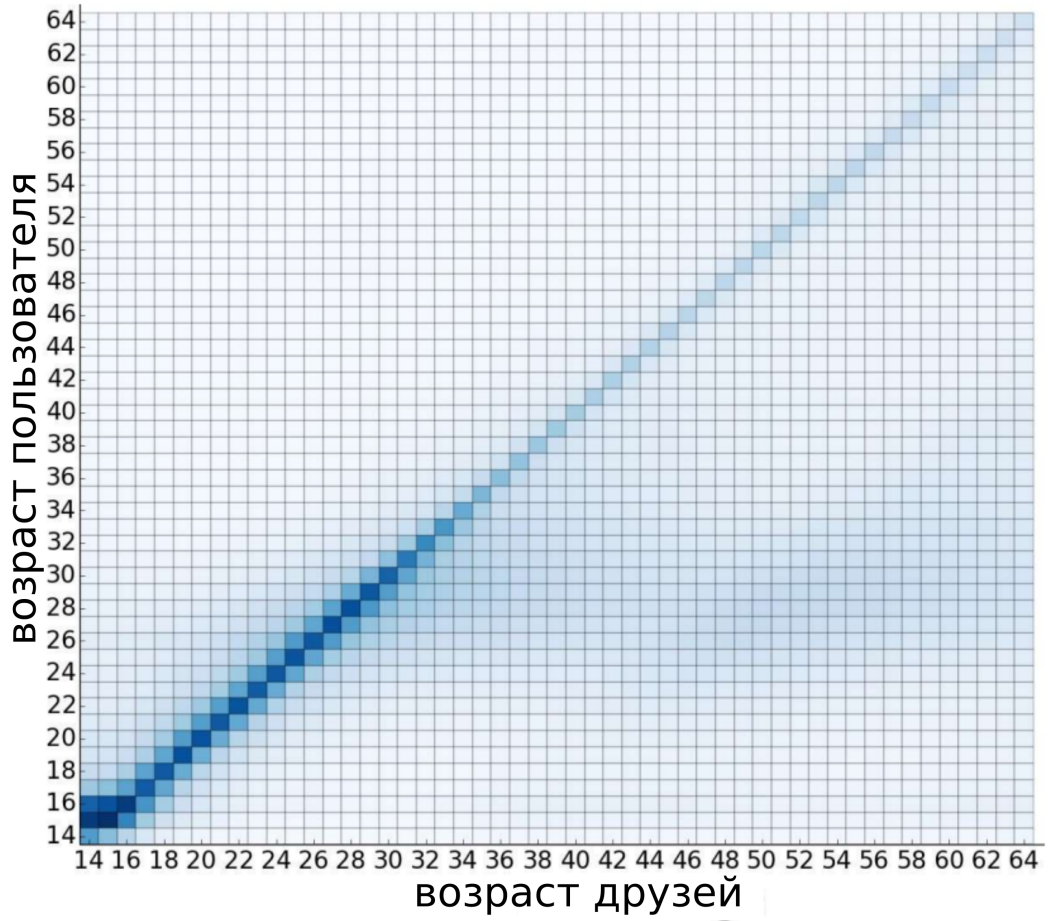

Рисунок 1.1 — Распределения возрастов соседних вершин в социальной сети Вконтакте

список друзей находится в открытом доступе, а у другого закрыт, в статистику попадает только одна точка (возраст, возраст друга).

Стоит отметить, что «гомофилия» может наблюдаться не только для демографических атрибутов, но и для структурных свойств. Так, например, в работе [53] на примере социальной сети Facebook было показано, что количество друзей для пользователя коррелирует со средним количеством друзей среди его друзей.

В общем случае наиболее популярные решения задачи предсказания значений демографических атрибутов по социальному графу можно разделить на три подхода. Первый подход заключается в применении методов машинного обучения без учителя. Задача предсказания значений характеристик пользователей сводится к задаче кластеризации вершин в графе. Методы кластеризации графа разбивают множество вершин на кластеры максимизируя количество связей (рёбер) внутри кластеров и минимизируя количество связей вершин из разных кластеров [54]. В контексте рассматриваемой в диссертации зада-

чи предполагается, что пользователи объединяются в кластеры согласно их демографическим характеристикам, т.е. кластер представляет собой множество пользователей с равными или похожими значениями рассматриваемо атрибута. Наиболее популярными методами кластеризации, являются алгоритмы распространения меток в графе [29; 55]. На каждом шаге алгоритма метка каждый вершины обновляется, используя метки соседних вершин. В нашем случае метка представляет собой некоторое значение рассматриваемого атрибута.

Второй подход заключается в применении методов машинного обучения с учителем. Идея заключается в построении векторных представлений вершин социального графа, представляющих пользователей [56; 57]. Представления вершин строятся только с использованием информации о структуре графа, т.е. связях между вершинами. Будем называть такие представления статическими. Затем данные вектора используются в качестве признаков для алгоритмов машинного обучения с учителем. Алгоритм обучается на признаках и значениях атрибута для тех пользователей, у которых это значение известно. Обученная модель используется для предсказания значений по признакам для пользователей с неизвестным значением атрибута.

Современные методы основаны на третьем подходе, заключающимся в использовании графовых нейронных сетей [58] (GNN, англ. Graph Neural Networks). Векторные представления для вершин, которые представляющие слои нейронной сети, учитывают не только связи между вершинами, но информацию о вершинах (например, векторные представления текстов пользователей) и известные значения предсказываемого атрибута. Таким образом, векторные представления строятся под конкретную задачу. Одним из примеров GNN, используемых для классификации, являются свёрточные графовые нейронные сети [59].

1.2.1 Методы на основе кластеризации графа

Одними из простых и популярных методов кластеризации графов являются методы, основанные на распространении меток в графе [29; 55]. Распространение меток – класс алгоритмов для обработки графов. Это итеративный процесс, на каждом шаге которого метки вершин (кластеры, значения

атрибутов и т.д.) обновляются для каждой вершины используя метки соседних для заданной вершины вершин. Алгоритмы распространения меток можно разделить на асинхронные и синхронные. В асинхронной версии на каждой итерации определяется порядок обхода вершин, в котором обновляются метки узлов. Вершина может получить как уже обновленные на данной итерации метки соседних вершин, так и метки, полученные на предыдущей итерации [29]. Схема данного подхода представлена в алгоритме 1. В синхронной версии метки всех вершин на каждом шаге вычисляются, используя только метки соседей, полученные на предыдущем шаге и обновляются синхронно между итерациями [60]. Схема данного подхода представлена в алгоритме 2. Преимущество синхронной версии в том, что такой алгоритм легко реализовать параллельно, так как нет зависимостей между метками вершин в рамках одной итерации.

Вход : $G = (V,E)$ – граф, L – множество меток
Выход: $Y : V \to L$ – метки вершин

```
1 init(Y); // инициализировать метки вершин
2 repeat
3     // проход по вершинам в определённом порядке
4     for x ∈ shuffle(V) do
5         N_1 := getLinks(x); // получить множество инцидентных
           вершин
6         Y[x] :=select({Y[z] : z ∈ N_1}); // выбрать метку среди меток
           соседей
7     end
8 until criteria(...);
9 return Y;
```

Алгоритм 1: Схема работы асинхронной версии алгоритма распространения меток

Работы, посвященные предсказанию демографических характеристик пользователей в социальном графе с помощью алгоритма распространения меток, используют синхронную версию алгоритма. Далее описываются работы, использующие алгоритм распространения меток и другие методы кластеризации графа для предсказания демографических характеристик пользователей.

Юргенс [61] применял алгоритм распространения меток для предсказания основного местоположения пользователей социальной сети. Метка представля-

Вход : $G = (V,E)$ – граф, L – множество меток
Выход: $Y : V \to L$ – метки вершин

```
1  init(Y); // инициализировать метки вершин
2  repeat
3      // проход по вершинам в произвольном порядке
4      for x ∈ V do
5          N_1 := getLinks(x); // получить множество инцидентных
            вершин
6          Y_0[x] :=select({Y_0[z] : z ∈ N_1}); // выбрать метку среди
            меток соседей
7      end
8      Y := Y_0; // обновить состояние целиком
9  until criteria(...);
10 return Y;
```

Алгоритм 2: Схема работы синхронной версии алгоритма распространения меток

ет собой географические координаты. Алгоритм итеративно вычисляет метку для каждого пользователя. Метка выбирается по меткам его соседей в графе (друзей). В работе рассматривалось несколько способов выбора метки по множеству меток соседей, специфичных для задачи предсказания местоположения.

Лиао и др. [62] предлагают метод предсказания возраста с использованием социального графа. Идея метода основана на эмпирическом наблюдении, что чем больше размер максимальной клики в графе, тем меньше разница возраста пользователей, входящих в клику. Другими словами, чем больше группа пользователей, где все связаны друг с другом, тем более вероятно, что эти пользователи из одной узкой возрастной группы. В основе предлагаемого авторами метода лежит алгоритм распространения меток. Метка вершины выбирается с учётом веса ребра. Вес ребра между двумя вершинами определяется как наибольший размер максимальной клики в графе, содержащей обе вершины. Для поиска максимальных клик в графе используется алгоритм Брона-Кербоша [63]. Для оценки качества метода авторы использовали два набора данных: из социальной сети MG и Twiter. Референсные значения возраста для первого набора извлекались из профиля пользователя. Для социальной сети Twitter использовались эвристики для поиска явного указания возраста в поле «о се-

бе» профиля, а также в сообщения с упоминанием пользователя и с текстом, содержащем поздравление с X-летием.

Ли и др. [64] рассматривают задачу предсказания значений атрибутов пользователей в следующей постановке. Рассматривается пользователь, у которого неизвестно значение атрибута. Для данного пользователя известна первая окрестность, или эго-сеть. Эго-сеть представляет собой подмножество социального графа, включающее в себя вершины, представляющие данного пользователя и связанных с ним пользователей. Эго-сеть включает в себя рёбра, представляющие все известные связи, включая связи данного пользователя, а также связи между пользователями из его первой окрестности. Для некоторых пользователей из данного подмножества графа известно значение рассматриваемого атрибута. Задача – предсказать значение атрибута для данного пользователя по эго-сети. Решение основано на моделировании так называемых кругов, представляющих локальные кластеры друзей в эго-сети заданного пользователя. В формальной модели, предложенной авторами, моделируются значения атрибута пользователей, круги, и связи между кругами. Зависимости определяются минимизируемой функцией стоимости. Эти зависимости предполагают, что пользователи из одного круга скорее имеют одинаковое значение атрибута, а также что пользователи из одного круга чаще связаны связаны между собой, чем с пользователями из других кругов. Процесс оптимизации представляет итеративный процесс, который может быть интерпретирован как распространение меток в эго-сети.

Доугнон и др. [65] используют подход, основанный на распространении меток в социальном графе для предсказания значений атрибутов пользователей. Особенностью задачи, поставленной в данном исследовании, является ограничение на максимальное количество вершин графа, используемых для предсказания значения атрибута для вершины. Это ограничение возникает в случае, если недоступен социальный граф всеё социальной, а для получения информации о вершине необходимы определённые ресурсы. При этом предполагается, что данные собираются и граф расширяется по мере необходимости при решении задачи. Еще одной особенностью является использование вершин различных типов. Помимо пользователей рассматриваются вершины, представляющие группы, а также вершины, соответствующие факту просмотра или отметки «мне нравится».

Филиппова [66] рассматривала задачу предсказания пола пользователей Youtube[5] по текстам комментариев. Рассматривается социальная сеть, представленная в виде двудольного графа (пользователи и видеоролики представлены узлами, просмотр видеоролика пользователем представлен ребром). В этом двудольном социальном графе распространяются метки – значений пола. Процесс распространения состоит из двух шагов: от пользователей к видеороликам и обратно. Автор использует известные и полученные с помощью распространения метки для улучшения качества предсказания пола пользователей по текстам их комментариев. Использование предсказанных с помощью распространения в графе значений пола позволяет увеличить обучающую выборку, а также исключить потенциальные ложные метки (случаи, когда распространенная метка не совпадает с исходной).

Подход, основанный на кластеризации графа, применяется и для других прикладных задач классификации. Например, Спериосу и др. [67] рассматривали задачу предсказания эмоциональной окраски публичных сообщений в социальной сети Twitter. Авторы строили граф, включающий себя узлы разных типов: пользователей, публичные сообщения, эмотиконы, хэштэги, n-граммы слов. Рёбра представляют отношения между вершинами: пользователь подписан на другого пользователя, пользователь является автором сообщения, эмотикон или n-грамма присутствует в сообщении и т.д. В качестве алгоритма кластеризации авторы применяли алгоритм MAD (англ. Modified adsorption), предложенный Талукдаром и Краммером [68]. Метки, представляющие значение эмоциональной окраски изначально инициализировались для части вершин, таких как эмотиконы, слова, n-граммы.

Мислове и др. [69] предсказывали значения атрибутов профиля студентов вуза, используя социальные связи. Рассматривались следующие атрибуты: колледж, основная специализация, отделение, школа, год поступления, родной город, политические предпочтения. Применялся итеративный метод кластеризации графа, предложенный Клаусетом [70]. На этапе инициализации пользователи с одинаковым значением атрибута попадали в один кластер. Для оценки качества предложенного метода авторы использовали два набора данных.

[5] https://youtube.com

1.2.2 Методы, основанные на статических векторных представлениях вершин графа

При решении задачи предсказания значений атрибутов пользователей по социальному графу с использованием алгоритмов обучения с учителем ключевым вопросом является представление вершин графа в пространстве признаков.

В работе Рао и др. [30], рассмотренной в разделе 1.1.3, помимо текстов сообщений использовалась информация о структуре социальной сети. В качестве признаков использовались количество подписчиков пользователя (входящие связи), количество пользователей, на которых подписан данный пользователь (исходящие связи), отношение количества входящих связей к количеству исходящих для каждого пользователя. Однако авторы отмечают, что использование эти признаков не показало приемлемого качества.

Дей и др. [71] рассматривали задачу предсказания возраста пользователей. Основываясь на свойстве «гомофилии», возраст пользователя оценивался с использованием известных значений возраста соседей. В качестве признаков рассматривалось среднее значение возраста среди соседей в графе, медианное значение возраста среди соседей, среднеквадратическое отклонение возраста среди соседей. На этих признаках обучалась модель линейной регрессии.

Хан и др. [72] также предсказывали значения атрибутов вершин графа. Авторы рассматривали набор паттернов, которые могу возникать в окрестности вершины графа и используют их в качестве признаков для вершины. Авторы проводят экспериментальную оценку предложенного подхода для графов, представляющие химическую структуру веществ, однако отмечают, что данный подход применим и для социальных графов.

Алзахрани и др. [73] выявляли среди аккаунтов Twitter такие, которые представляют организации. Для извлечения признаков из структуры социальной сети, авторы рассматривали три графа: цитирования, подписки, упоминания. Во всех графах вершины представляют аккаунты пользователей, некоторые из которых могут представлять организации. Рёбра между вершинами возникают, если пользователь цитировал сообщения другого пользователя, пользователь подписан на другого пользователя, пользователь упоминал другого пользователя в сообщениях. Для каждой вершины в каждом графе вычислялись несколько различных центральностей: степень вершины, PageRank [74],

K-core [75], коэффициент кластеризации [76]. Авторы использовали эти значения вместе с другими признаками, извлечёнными из профилей и метаданных твитов, для классификации аккаунтов. Рассматривались несколько классификаторов, наилучшим образом показала себя логистичекая регрессия.

Одним из подходов к получению векторного представления вершин является преобразование матриц, описывающих связи между вершинами. Примером такой матрицы матрицы является матрица смежности $A = (a_{ij})_{|V|\times|V|}$, которая для графа $G = (V,E)$ определяется как:

$$a_{ij} = \begin{cases} 1 & \text{если} (i,j) \in E \\ 0 & \text{иначе} \end{cases} \tag{1.2}$$

Особенностью подобных матриц является большая размерность и, в некоторых случаях, разреженность, что затрудняет её использование вместе с алгоритмами машинного обучения с учителем. Для уменьшения размерности применяются методы главных компонентов (англ. principal component analysis, PCA). Решения, основанные на использовании метода главных компонент, применяются в том числе и для предсказания атрибутов вершин в социальных графах.

Косински и др. [77] рассматривали множество атрибутов пользователей социальной сети Facebook, такие как сексуальная ориентация, политические взгляды, исповедуемая религия, удовлетворенность в жизни, возраст, пол, статус отношений и другие. Задача заключается в предсказании неизвестных значений атрибутов по отметкам «мне нравится» (лайкам) пользователей. Пользователи и их отметки «мне нравится» представляются в виде разреженной матрицы пользователь-лайк, в которой элемент равен 1, если пользователь поставил отметку «мне нравится» определенному объекту. Полученная матрица является матрицей смежности для двудольного графа пользователь-лайк. Размерность этой матрицы уменьшалась с использованием метода на основе сингулярного разложения (англ. SVD – singular value decomposition). Затем уже плотная матрица уменьшенной размерности использовалась в качестве входных данных для алгоритмов машинного обучения с учителем. В качестве таких алгоритмов авторы рассматривали логистическую или линейную регрессию, в зависимости от атрибута.

Бызов, Губанов и др. [78] рассматривали задачу предсказания политических предпочтений пользователей социальной сети Вконтакте. Рассматрива-

лось три класса политических взглядов: Социалист, Либерал, Державник. Для предсказания использовались социальные связи пользователей, включающие дружбу между пользователями и подписки пользователей на сообщества. С целью уменьшения размерности признакового пространства использовался метод на основе SVD разложения матрицы смежности социального графа. К преобразованным признакам применялась статистическая модель логистической регрессии. Средняя полнота на наборе данных, собранном из социальной сети составила более 67%. Авторы применяют предсказанные политические предпочтения для нахождения оптимальной позиции политика или политической партии с целью формирования стратегии идеологического позиционирования.

Хуанг и др. [79] использовали информацию из профилей пользователей, тексты сообщений и социальные связи для предсказания рода деятельности социальной сети Weibo. В качестве признаков, извлекаемых из социального графа авторы использовали скрытую структуру сообществ. Поиск скрытой структура в ненаправленном графе $G = (V,E)$ производился таким способом, чтобы максимизировать модулярность [80]:

$$Q = \frac{1}{2|V|} \sum_{i,j} \left[a_{ij} - \frac{d_i d_j}{2|E|}\right] \mathbb{1}(g_i = g_j) \tag{1.3}$$

Здесь $A = (a_{ij})_{|V|\times|V|}$ – матрица смежности графа (1.2), $d_i = \sum_k a_{ik}$ – степень вершины i, g_i – сообщество вершины i.

Авторы рассматривали матрицу модулярности $B = (b_{ij})_{|V|\times|V|}$, которая определяется как:

$$b_{ij} = a_{ij} - \frac{d_i d_j}{2|E|} \tag{1.4}$$

Для поиска нечетких сообществ в графе авторы производили разложение матрицы $B = UDU^T$, где $U = (u_1|u_2|\dots)$ – матрица, составленная из собственных векторов B, D – диагональная матрица, составленная из собственных значений β_j матрицы B. Затем задача сводилась к максимизации значения модулярности, которое выражается как:

$$Q = \frac{1}{2|E|} \sum_{j=1}^{|V|} \sum_{k=1}^{c} \beta_j (u_j^T s_k)^2 \tag{1.5}$$

Здесь Q – значение модулярности, c – количество сообществ, s_k – вектор, описывающий k-е сообщество, т.е. $s_{ki} \in [0,1]$ означает степень принадлежности

пользователя i сообществу k. Полученные нечёткие сообщества использовались в качестве признаков для предсказания рода деятельности с использованием классификаторов SVM и логистической регрессии. Для оценки качества метода в использовались референсные значения рода деятельности для пользователей, полученные из списка верифицированных аккаунтов Weibo.

В работе [81] предложена модель DeepWalk, которая использует случайные блуждания для получения векторного представления вершин графа. Алгоритм построения векторов для вершин состоит из двух этапов:

1. Случайные блуждания по графу порождают последовательности вершин.
2. На основе полученных последовательностей вершин генерируется векторное представление вершин. На данном этапе применяется неглубокая нейронная сеть, по аналогии с моделью word2vec [38].

Пероззи и др. [82] рассматривали задачу предсказания возраста пользователей социальной сети. В отличие от других работ (например, [31; 83]) авторы не разбивали значения возраста, чтобы свести задачу к классификации. Вместо этого авторы предсказывали значение возраста с точностью до года. Таким образом, задача сводилась к регрессии. В качестве векторных представлений (вложений) вершин социального графа авторы рассматривали DeepWalk [81]. Для обучения и предсказания значений возраста по полученным признакам применяется линейная регрессия. Авторы использовали набор данных Pokec для экспериментального исследования предложенного метода. Этот набор данных собран и опубликован Такацом и др. [84]. Набор представляет собой дамп открытых данных социальной сети Pokec, включающий социальный граф и профили пользователей.

Алетрас и др. [85] использовали тесты пользователей и социальный граф для предсказания рода деятельности и дохода пользователей. Авторы расширили своё предыдущее решение [35], использующее только текстовые данные, признаками, построенными по социальным связям. Часть набора данных, включающего в себя социальный граф и значения рассматриваемых атрибутов находится в отрытом доступе, что позволяет использовать его для оценки методов в рамках диссертационной работы. Авторы использовали метод DeepWalk [81] для построения векторных представлений (вложений) вершин графа. К полученным представлениям применялись различные алгоритмы машинного обучения: линейная регрессия, логистическая регессия, гауссовский

процесс, SVM. В работе авторы уделяли особое внимание настройке алгоритмов, т.е. подбору гиперпараметров. Таким образом метод решения исходной задачи заключается в следующем: на обучающем наборе производится перекрестная проверка, в которой выбирается наилучшие значения гипер-параметров для заданного классификатора/регрессора, затем настроенный с выбранными гипер-параметрами алгоритм применятся для предсказания неизвестных значений атрибута. Обычно для оценки качества метода, в котором не производится подбор гипер-параметров применяется перекрестная проверка. Так как в данном случае классификаторы настраиваются, для оценки качества предлагаемого подхода авторы применяли вложенную перекрёстную проверку. В работе представлены результаты вложенной перекрёстной проверки как отдельно для методов, использующих только текстовые данные и только социальный граф, так и для методов, использующих конкатенацию текстовых и графовых признаков (векторных представлений вершин графа).

Иванов и др. [86] предложили подход получения векторного представления вершин графа BLM, который основан на вероятностной модели рёбер. В данном подходе каждая вершина представлена двумя векторами, $In \in R^D$ и $Out \in R^D$ (где D – размерность векторного пространства), которые являются векторным представлением входящих и исходящих рёбер этой вершины соответственно. Вероятность ребра между вершинами вычисляется по формуле:

$$p(v|u) = \frac{\exp(In_u^T Out_v)}{\sum_{w \in V} \exp(In_u^T Out_w)} \tag{1.6}$$

Векторное представление может быть получено с помощью принципа максимума правдоподобия, однако непосредственное получение векторного представления затруднено необходимостью вычислять сумму в знаменателе формулы (1.6). Вместо этого авторы работы используют метод контрастной оптимизации [87].

Трофимович и др. [88] использовали векторные представления BLM для предсказания местоположения пользователей социальной сети. Так как местоположение предсказывалось для социального графа, в котором связи ненаправленные, авторы полагали $In_u \equiv Out_v$. В качестве классификаторов рассматривались логистическая регрессия, случайный лес, XGboost, а также многослойные нейронные сети. Результаты работы системы были оценены с помощью кросс-валидации. Наилучшие результаты показала многослойная нейронная сеть.

1.2.3 Методы, основанные на графовых нейронных сетях

Графовые нейронные сети [58] – современный фреймворк, предназначенный для решения прикладных задач на графах. Основным его преимуществом является то, что при обучении модели учитываются не только связи между вершинами, то есть граф, но и информация о вершинах и рёбрах: признаковые представления вершин, метки рёбер. Одной из разновидностей графовых нейронных сетей, применяемых на практике, является свёрточные графовые нейронные сети. Киф и Веллинг [59] применяли её для решения задач классификации на графах. Авторы проводили экспериментальное исследование графовых нейронных сетей на графах цитирования документов и графах, представляющих базу знаний. Во всех рассматриваемых наборах данных у каждой вершины имелся вектор признаков, являющийся некоторым представлением моделируемого этой вершиной объекта.

Мак Ким и др. [89] представили фреймворк, который позволяет эффективно сочетать граф социальной сети, тексты пользователей и значения меток для предсказания демографических атрибутов пользователей Twitter. Фреймворк реализует модель графовых нейронных сетей. Для решения задачи предсказания значений атрибутов авторы строили рекурсивную нейронную сеть, структура которой соответствует структуре социальной сети. Для каждого пользователя строится дерево из рекурсивных нейронных единиц согласно структуре окрестности данного пользователя в графе. Вершины, представляющие пользователей, становятся нейронными единицами, при этом заданный пользователь располагается в корне дерева. Рёбра графа определяют потоки данных между соответствующими нейронными единицами (от листьев к корню). Рёбра между узлами на k-й окрестности данного пользователя не используются, таким образом получается древовидная структура. Каждая нейронная единица принимает на вход признаки, извлечённые из текстов соответствующего пользователя, и выходы нейронных единиц, соответствующих соседям вершины в социальном графе. В качестве нейронных единиц авторы используют два вида единиц: наивную и основанную на LSTM [51; 90]. Полученную нейронную сеть авторы использовали для предсказания неизвестных значений пола, возраста, типа пользователя социальной сети Twitter.

Методы, основанные на применении графовых нейронных сетей для задач классификации, описанные в работах [59] и [89] предполагают наличие признаков у вершины, например, извлечённых из текстов. Однако эти методы не применимы для задач, где имеется лишь граф и отсутствует дополнительная информация о вершинах. Авторы фреймворка DGL (англ. Deep Graph Library) [91], предназначенного для решения прикладных задач с помощью графовых нейронных сетей, на одной из страниц документации приводят пример[6], в котором рассматривается задача классификации вершин. В обозначенной в примере задаче вершины не имеют признаков. Поэтому в качестве входных векторов для свёртчной нейронной сети используются обучаемые векторные представления вершин. Таким образом, при обучении метод позволит получить на выходе не только модель для предсказания класса для каждой вершины графа, но и векторные представления вершин, построенные под конкретную задачу классификации.

Автором диссертации не было найдено научных работ, применяющих описанный подход для предсказания значений демографических атрибутов по социальному графу. Однако, как будет показано в главе 3, он показывает наилучшее качество при классификации вершин по значениям демографических атрибутов на используемых наборах данных.

1.3 Другие методы определения значений демографических атрибутов

Большинство описанных выше методов рассматривают задачу предсказания значений одного атрибута по одному виду данных: граф или текст. В данном подразделе уделяется внимание работам, в которых эти ограничения ослабляются. Сначала описываются исследования, использующие другие виды данных для предсказания значений атрибутов. Затем рассматриваются методы совместного предсказания значений нескольких атрибутов. В конце обсуждаются способы комбинирования данных различной природы для решения задачи предсказания значений демографических атрибутов.

[6] https://docs.dgl.ai/tutorials/basics/1_first.html

Одним из признаков для определения пола пользователя является его имя. Танг и др. [92] изучали имена американских пользователей социальной сети Facebook. Так как в этой социальной сети чаще всего пользователи указывают своё настоящее имя, то для предсказания пола применялся простой метод, основанный на использовании списков имён. Помимо словаря имён новорожденных детей США авторы использовали свой словарь, который получили из собранных профилей пользователей Facebook. Для каждого имени авторы эмпирически оценивали пол на основе информации из тех профилей, где указано значение пола.

Лиу и др. [93] изучали тот же вопрос для социальной сети Twitter. По оценкам авторов, для 66% пользователей метод, основанный на определении пола по имени с помощью словарей имён, не даёт верного результата, так как в имя часто представляет собой никнейм, аббривеатуру, либо иной текст, не являющийся настоящим именем человека. Для предсказания пола авторы использовали тексты сообщений, извлекая из них признаки. Однако было показано, что использование имён, для которых имеется статистика о поле пользователей с данными именем, позволяет улучшить качество предсказания значений данного атрибута.

Аловибди и др. [94] предложили подход для предсказания пола пользователей Twitter по имени. Его преимущество заключается в том, что решение не зависит от языка, на котором написано имя. В качестве признаков для предсказания авторы использовали фонемы, составляющие имя. В качестве алгоритмов для классификации использовались наивный байесовский классификатор (NB), дерево принятия решений (DT) и комбинированный метод NB-Tree, представляющий дерево принятия решений, в листьях которого располагается наивный байесовский классификатор.

Маккорристон и др. [95] рассматривали одну из задач предсказания типа аккаунта в социальной сети Twitter. Авторы выделяли аккаунты, представляющие организации, и персональные аккаунты пользователей. Использовался подход, основанный на на машинном обучении с учителем. В качестве признаков использовалась извлечённая из текстов сообщений и из социальных связей метаинформация: слова, наиболее специфичные для каждого из классов, средняя длина слова, частота хэштэгов и гиперссылок, доля ретвитов среди всех твитов, отношение количества входящих рёбер (подписчиков) к количеству

исходящих рёбер. В качестве классификатора использовался метод опорных векторов (SVM).

Донг и др. [83] рассматривали задачу совместного предсказания значений двух атрибутов: пола и возраста. Авторы проанализировали большой набор данных, включающий в себя взаимодействие абонентов сотовой связи, представленный в виде графа. Абоненты представлены узлами, пары абонентов, которые взаимодействовали посредством сотовой связи, представлены рёбрами. В рамках анализа графа и значений пола и возраста абонентов авторы сделали несколько наблюдений. Например, люди чаще взаимодействуют с пользователями обоих полов в репродуктивно-активном возрасте, в периоде знакомств. Однако наблюдается тенденция преимущественного взаимодействия с людьми того же пола начиная со зрелого возраста. Данные наблюдения легли в основу метода, в рамках которого задача предсказания пола и возраста формализована в виде вероятностной модели для классификации двух зависимых переменных. Таким образом, пол и возраст предсказываются одновременно.

Хуанг и др. [79] рассматривали задачу совместного предсказания значений нескольких атрибутов пользователей: пола, статуса семейных отношений, уровня образования. Рассматривался двудольный граф пользователь-товар. Ребро в графе означает, что пользователь покупал товар. Метод основан на нейронных сетях. Входной слой представляет собой разреженный вектор, представляющий для каждого пользователя множество купленных им товаров. Далее в нейронной сети представлен общий скрытый слой, затем слой, представляющий специфичные для каждого атрибута скрытые представления, к которым применяется Softmax. Функция перехода от входного слоя к общему обучается так, чтобы минимизировать ошибку для всех предсказываемых атрибутов. Таким образом, модель позволяет учитывает возможную зависимость между значениями различных атрибутов.

Ал Замал и др. [96] предсказывали пол, возраст, политические взгляды для пользователей социальной сети Twitter. Для этого использовались два вида данных: тексты пользователей и социальный граф. Авторы предложили извлекать текстовые признаки не только для пользователей, для которых делается предсказание, но и для их друзей. Экспериментальная оценка метода показала, что использование признаков, извлеченных из текстов соседних пользователей в графе, вместе с признаками, извлеченными из текстов непосредственно це-

левого пользователя, улучшает качество предсказания значений атрибутов по сравнению с использованием только текстов целевого пользователя.

Работы Алетрас, Преотиуц-Пиетро и др [35; 85] уже были описаны в разделах 1.1 и 1.2. В качестве текстовых признаков использовались кластеры слов, вычисленные с использованием представлений word2vec [38]. В качестве представлений пользователей, извлечённых из структуры социального графа, авторы используют модель DeepWalk [81]. Затем признаки, извлечённые из текстов конкатенировались с признаками, полученными из структуры графа. К полученным представлениям применялись метод опорных векторов, гауссовский процесс, логистическая и линейная регрессия для предсказания рода деятельности пользователей и их дохода. Комбинированные признаки улучшили метод, который использовал только текстовые признаки.

Еще одним из способов комбинировать тексты пользователей и их социальные связи, является использование графовых нейронных сетей. Такой подход используется, например, в работе Мак Ким и др. [89], описанной в разделе 1.2.

Гнатышак, Игнатов, Кузнецов и др. [97] рассматривали задачу выявления групп пользователей со схожими интересами и сообщества пользователей принадлежащих сходным группам. Под группой понимается специальный тип страницы в социальной сети Вконтакте, на которую подписываются пользователи. Как правило, группа объединяет пользователей по интересам, которые изначально не всегда заданы явно. В представленной в работе модели рассматриваются три типа сущностей в социальной сети: пользователи, группы и интересы. Предложен метод поиска плотных трикластеров вида (Users, Groups, Interests). Каждый трикластер описывает множество интересов, множество групп по этим интересам и множество пользователей по этим интересам, которые подписаны на эти группы. В качестве возможного применения трикластеров авторы отмечают формирование рекомендаций для пользователя других потенциально интересных пользователей, групп и ресурсов, с учётом интересов.

1.4 Особенности сбора данных и оценки качества методов

Одним из этапов при построении решений задач определения значений демографических атрибутов является сбор данных. Кроме того, данные необ-

ходимы для сравнения качества различных методов. В случае предсказания значений демографических атрибутов по текстам сообщений пользователей набор данных содержит информацию о множестве пользователей. Для каждого пользователя информация включает в себя тексты сообщений, написанных этим пользователем и значения демографических атрибутов. Для задач предсказания значений атрибутов пользователей по социальным связям набор дланных включает в себя социальный граф и значения демографических атрибутов для некоторых пользователей, представленных вершинами в этом социального графа.

Референсные значения атрибутов пользователей. Для получения референсных значений демографических атрибутов используется несколько подходов. Одним из подходов основан на использовании сервиса поиска пользователей и друзей в социальных сетях. Например, в социальной сети Вконтакте сервис поиска пользователей позволяет явно указать значения атрибутов интересующих пользователей. Стоит отметить, что такой подход не является таргетированным сбором для заданного множества пользователей. Второй подход заключается в ручной разметке. Значения целевых атрибутов определяются экспертами. Подход применялся в работах [28; 30; 33; 34]. Профили пользователей в анализируемой социальной сети могут не содержать нужный атрибут или значения нужного атрибута. Примером такого ресурса является Twitter. В профиле Twitter нет полей для указания пола, возраста, семейного положения и других атрибутов. Значение нужного атрибута может быть указано пользователем в других социальных сетях. Задача заключается в том, чтобы найти профили пользователя на других ресурсах. В профиле Twitter имеется поле URL, в котором пользователи часто указывают гиперссылку на свой профиль в другой социальной сети. Подход, при котором референсные значения извлекаются из других ресурсов, используется в работах [26; 94].

Оценка качества методов предсказания значений атрибутов. В общем случае задачи предсказания значений атрибутов сводятся к задачам классификации или регрессии, в зависимости от атрибута и постановки задачи. Например, задача предсказания значений пола пользователей (мужской или женский) сводится к задаче классификации. Задача предсказания возраста

может быть сведена как к классификации, так и к регрессии. Первый случай возникает, например, при разбиении значений возраста на интервалы, и в задаче важно верно предсказать возрастной интервал. Если предсказывается точное значение возраста, и ошибка в 1-2 года не так критична, как ошибка в 20 лет, то такая задача сводится к задаче регрессии. При оценке качества методов в работах используются традиционные метрики. Достоверность, F-1 мера с микро- и макроусреднением используются для оценки качества методов, рассматривающих задачу предсказания значений атрибутов как задачу классификации. При оценке качества методов, рассматривающих исходную задачу к задаче регрессии, используются среднеквадратичная ошибка (MAE), коэффициент детерминации (R2), коэффициент Пирсона.

Сэмплинг. При сборе связного социального графа для части социальной сети, а также с целью получения репрезентативной выборки пользователей, возникает задача обхода вершин графа. Процесс обхода вершин называют также сэмплингом. Исследования [98; 99] показали, что наиболее репрезентативные выборки получаются при использовании алгоритмов сэмплирования Forest Fire и Метрополиса-Гастингса.

В контексте задачи определения значений демографических атрибутов для заданного множества пользователей часть социальной сети собирается таргетированно. Это означает, что собранный граф содержит информацию об окружении вершин, соответствующих целевым пользователям. В работе [85] граф собирается следующим образом. Для каждого целевого пользователя собирается множество идентификаторов страниц, на которые он подписан. Затем формируется граф, включающий вершины, соответствующие целевым пользователям и вершины, связанные с целевыми вершинами. В граф включены такие рёбра, для которых хотя бы одна из инцидентных вершин соответствует целевому пользователю.

Автором диссертационной работы было проведено экспериментальное исследование, цель которого понять, достаточно ли собранных связей для предсказания значений рода деятельности пользователей, или необходима информация о всех возможных связях между всеми вершинами в графе.

Формализуем задачу. Пусть $\mathbf{L}$ – множество возможных значений атрибута, $U = \{v_1, \ldots, v_n\}$ – множество целевых пользователей, для t из них ($t < n$) известно значение атрибута. Обозначим эти значения как $Y_t = [y_1, \ldots, y_t]$, где

$y_i \in \mathbf{L}$ – деятельность пользователя v_i. Значения атрибута для $v_{t+1}, \ldots, v_n$ неизвестны и будут предсказываться. Обозначим их как $Y_p = [y_{t+1}, \ldots, y_n]$. Рассмотрим социальный граф $G = (V, E)$. Вершины $V = \{v_1, \ldots, v_m\}$ представляют пользователей U и другие страницы, на которые подписаны U $(m > n)$, направленные рёбра $E = \{x \to y | x \in U, y \in V\}$ представляют отношения подписки целевых пользователей на вершины V. Рассмотрим также расширенную версию графа $G^* = (V, E^*)$, где $E^* = \{x \to y \mid x \in V, y \in V\}$ – все известные рёбра между вершинами V. Задача – предсказать Y_p при заданном графе (G или G^*) и Y_t.

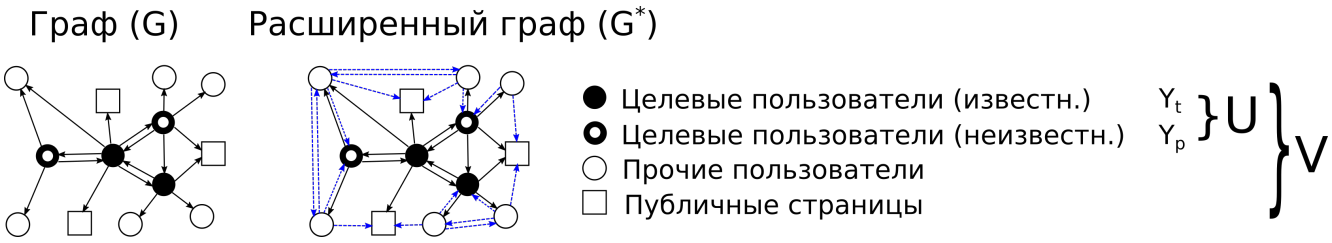

Рисунок 1.2 — Социальный граф и расширенный социальный граф

Экспериментальное сравнение проводилось для собранного набора данных из социальной сети Вконтакте. Значения рода деятельности для пользователей собирались с помощью аннотаторов-добровольцев. Применяются методы, основанные на существующих решениях: алгоритм распространения меток, векторные представления вершин графа (LINE [100] и DeepWalk [81]) в качестве входа для классификатора XGBoost] [101]. Рассматриваются также признаки Distr, представляющие распределение значений рода деятельности среди соседей заданной вершины. Они также используются в качестве входа для классификатора XGBoost. Рассматривается также конкатенация признаков DeepWalk и Distr в качестве признакового представления пользователя.

В рамках эксперимента сравнивается качество предсказания рода деятельности пользователей для двух графов: G и его расширенной версии G^*. Стоит отметить, что сбор графа G^* требует значительно больше ресурсов по сравнению со сбором G. На рисунке 1.3 представлены значения и доверительные интервалы для значений F-меры, полученные в результате применения скользящего контроля. Алгоритм распространения меток обозначен как LP, векторные представления DeepWalk как DW, XGBoost как XGB. Из результатов следует, что для задачи предсказания рода деятельности дополнительные рёбра расширенного графа не улучшают качество предсказания на собранном наборе данных. Таким образом, для решения этой задачи достаточно собрать лишь

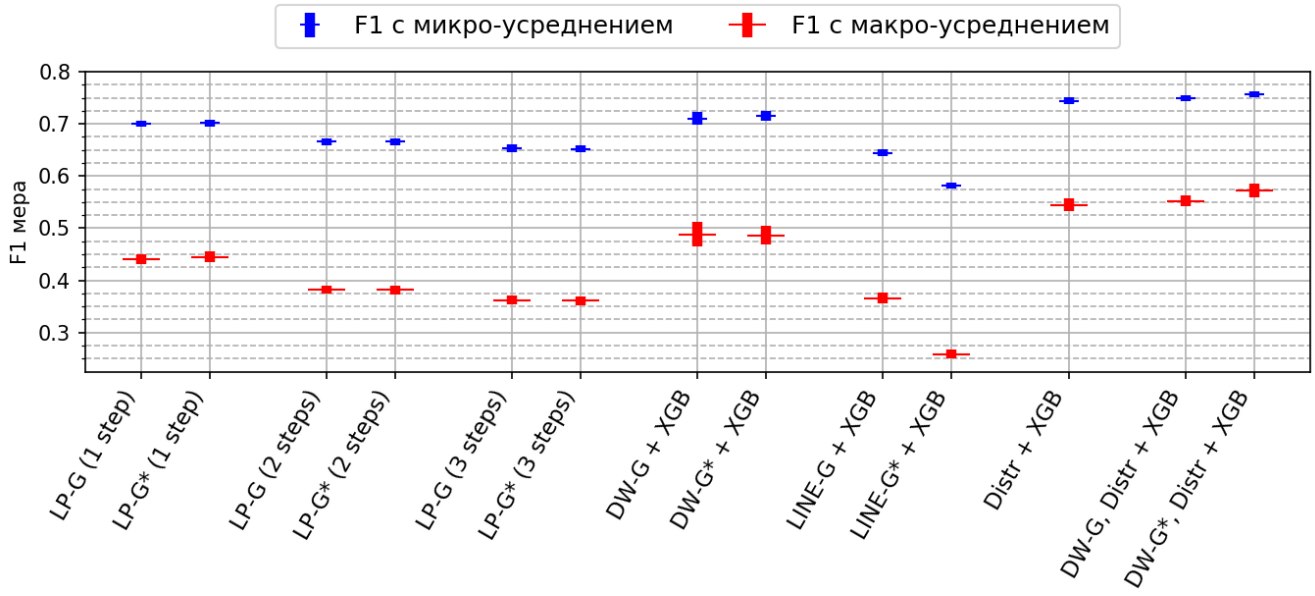

Рисунок 1.3 — Результаты оценки качества методов предсказания рода деятельности

граф G, как это сделано в работе [85]. Результаты опубликованы в виде тезисов [5] и доложены на конференции «Ломоносовские чтения» в 2020 году.

Ограничения при сборе и распространении данных. При сборе больших наборов данных также стоит обратить внимание на ограничения. Так, например, API Twitter позволяет выдавать ограниченное количество ответов на запросы к данным в течение заданного промежутка времени. Эти ограничения могут быть неявными, например, при слишком частом обращении к сервису, он может заблокировать как отдельные запросы, так и все последующие запросы пользователя. Кроме того, правила использования социальных сетей могут ограничивать распространение собранных данных. Это существенно усложняет процесс экспериментального сравнения различных методов предсказания значений демографических атрибутов.

1.5 Недостатки существующих методов

В конце раздела 1.1 были обозначены недостатки методов определения значений демографических атрибутов для заданного множества пользователей по текстам комментариев в социальной сети. Они заключаются в затруднённости

таргетированного сбора текстов комментариев для заданного пользователя и отсутствием публичных комментариев у многих пользователей. В общем случае даже в рамках одной социальной сети тексты разнородны. Они могут существенно отличаться по длине, орфографии, семантике. Например, тексты комментариев к сообщениям обычно короткие, содержат опечатки. Длинные публикации как правило описывают истории из жизни людей, эти тексты обычно выверены, написаны в определённом стиле, ошибки и опечатки исправлены до публикации. Такое разнообразие также затрудняет построение универсального решения на основе анализа текстов пользователей.

Методы предсказания значений демографических атрибутов по социальным связям не лишены недостатков. В частности, неверно работают в следующей ситуации, проиллюстрированной рисунком 1.4. Допустим, пользователь, являющийся разработчиком ПО (обозначен x), подписан на две страницы крупных сообществ, например, на новостной источник и страницу, посвящённую творчеству популярного артиста (a и c). Профессия подписчиков каждого из этих сообществ равномерно распределена, с небольшим преобладанием, например, водителей (красный цвет вершин). Кроме того, пользователь подписан на одно небольшое тематическое сообщество, подписчики которого являются преимущественно разработчиками ПО (b на рисунке). Алгоритм распространения меток, например, будет предполагать, что этот пользователь является водителем, так на первой итерации вершины a и c получат красную метку, а специализированное сообщество разработчиков b получит синюю метку, на второй итерации для вершины x будет выбрана красная метка. Для этого алгоритма подписка пользователя на популярную независимо от профессии страницу и подписка на узкоспециализированное профессиональное сообщество имеют одинаковую значимость.

Методы, описываемые в диссертационной работе, учитывают более высокую значимость узкоспециализированных сообществ и показывают более высокое качество по сравнению с другими методами.

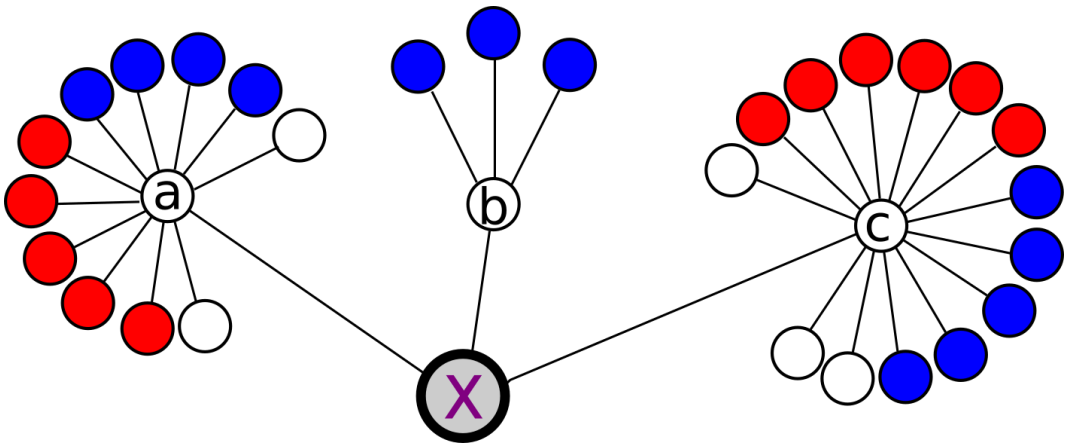

Рисунок 1.4 — Пример графа со специфичными и неспецифичными вершинами

1.6 Выводы

В первой главе был проведён обзор методов предсказания значений атрибутов пользователей социальных медиа.

Сначала были рассмотрены методы предсказания на основе анализа текстов пользователей. В обзор включены как ранние работы, посвященные анализу личных дневников, так и работы, описывающие методы предсказания значений атрибутов по текстам пользователей в микроблогах и других социальных медиа. Были выявлены недостатки методов предсказания значений демографических атрибутов по текстам сообщений в применении к постановке задаче с заданным набором целевых пользователей.

Далее были выделены подходы к предсказанию значений демографических атрибутов по социальному графу и рассмотрены методы на основе кластеризации графа, использования статических векторных представлений вершин графа и графовых нейронных сетей. Был обозначен недостаток методов, заключающийся в том, что алгоритмы не отличают значимость различных подписок пользователя на публичные страницы для определения значения его атрибута.

В главе также описаны методы совместного предсказания значений нескольких демографических атрибутов, способы комбинирования текстов пользователей и социального графа для предсказания значений атрибутов. Описаны методы, использующие отличные от текстов и социального графа данные.

Рассмотрены некоторые аспекты сбора данных. Описаны способы извлечения референсных значений, необходимых для обучения и оценки качества методов. Рассмотрены методы сэмплинга, то есть обхода вершин с целью получения репрезентативной выборки социальной сети. Описаны результаты экспериментального сравнения двух способов сбора социального графа для заданного множества целевых пользователей. Также перечислены используемые в работах способы оценки качества методов предсказания значений демографических атрибутов пользователей социальных медиа.

Обзор методов определения значений демографических атрибутов по текстам сообщений опубликован в статье [2]. Результаты экспериментального сравнения методов предсказания значений возраста и уровня образования пользователей социальной сети Вконтакте по текстам их публичных комментариев опубликован в статье [4]. Анализ различных сэмплов для предсказания значений рода деятельности пользователей Вконтакте опубликован в работе [5].

Глава 2. Подход для предсказания значений атрибутов пользователей на основе специфичности контекста

В главе рассматривается задача предсказания значений заданного демографического атрибута пользователей социальных сетей по социальному графу. Описываются наборы данных собранные из реальных социальных сетей, представляющие собой социальный граф и референсные значения атрибутов пользователей. Вводится новое понятие специфичности контекста и экспериментально, с использованием этих наборов данных, показывается, что специфичность контекста может служить хорошим признаком для предсказания значений атрибута пользователей в социальном графе. В конце главы описывается общий подход к предсказанию значений атрибутов пользователей на основе использования специфичности контекста.

2.1 Обозначения

$G = (V, E)$ – ненаправленный граф без петель с множеством вершин V и множеством рёбер E.

$N_x^1 = N_x = \{z \in V : (x, z) \in E\}$ – множество соседних вершин вершины x в графе $G(V, E)$. Также будем называть его *первой окрестностью*, *соседями*, *множеством соседей*.

$N_x^2 = \{z : (x, v) \in E, (v, z) \in E, z \neq x\}$ – *двухшаговая окрестность*, множество вершин, достижимых из x ровно в два шага (перехода по рёбрам), кроме самой вершины x.

$\mathcal{A}$ – рассматриваемый атрибут.

$A = (a_1, a_2, \ldots, a_{|A|})$ – возможные значения атрибута $\mathcal{A}$.

$U \subseteq V$ – множество вершин, представляющие пользователей в графе $G = (V, E)$.

$y_x \in A$ – значение атрибута $\mathcal{A}$ для вершины $x \in U$.

$Y = \{x \in U : \exists y_x\}$ – множество *размеченных пользователей*, т.е. вершин, у которых известно значение атрибута.

Определим функцию d: $2^V \rightarrow \mathbb{R}^{|A|}$, которая для заданного множество вершин X определяет вектор, представляющий дискретное распределение по значениям атрибута:

$$\begin{cases} d(X)_i = \dfrac{\left|\{x \in X \cap Y : y_x = a_i\}\right|}{\left|X \cap Y\right|}, & \text{если } X \cap Y \neq \varnothing \\ d(X) \text{ не определено} & \text{иначе.} \end{cases} \tag{2.1}$$

Если $d(X)$ существует, то сумма его элементов равна 1, порядок элементов соответствует порядку в A.

2.2 Постановка задачи

Рассмотрим задачу предсказания значений атрибута $\mathcal{A}$ пользователей социальной сети по заданному социальному графу. Социальный граф $G(V, E)$ содержит вершины V, представляющие пользователей и другие публичные страницы (организации, группы по интересам и т.д.). Социальные отношения, такие как дружба двух пользователей и подписка пользователя на публичные страницы, представлены рёбрами E. Для некоторого подмножества пользователей $Y \subseteq V$ известны значения $y_x \in A$ рассматриваемого атрибута $\mathcal{A}$. Задача заключается в предсказании неизвестных значений атрибута $\mathcal{A}$ для остальных пользователей $U \setminus Y$.

Социальный граф $G(V, E)$ рассматривается как ненаправленный граф. Если пользователь $x \in U$ подписан на другую публичную страницу (пользователя, организацию и т.д.) $z \in V$, то E будет содержать ненаправленное ребро $(x,z) \equiv (z,x)$.

2.3 Используемые наборы данных

Описываемые а разделе наборы данных используются для проверки гипотез «гомофилии», зависимости между размером общего контекста и значениями

атрибутов, зависимости между специфичностью общего контекста и значениями атрибутов. Наборы данных также используются для экспериментального сравнения методов предсказания значений демографических атрибутов по социальному графу, которое будет описано в главе 3

В работе используются как существующие наборы данных (twitter, pokec), так и собственные (vk1, vk2). Количество вершин, количество размеченных вершин, количество рёбер, количество различных значений рассматриваемого атрибута для каждого набора данных представлены в Таблице 3. Во всех наборах данных предполагается, что граф ненаправленный. Далее наборы данных описываются более детально.

В наборах данных vk1 и vk2 под атрибутом возраст понимается атрибут год рождения пользователя. Социальный граф представляет собой снимок части социальной сети в определённый момент времени. Для рассматриваемых методов не имеет значения, как называются и что означают метки (значения возраста, года рождения или другие значения), для некоторых методов имеет значение лишь абсолютная разность значений возраста или года рождения, которая в данном случае тождественна (абсолютная разность в возрасте равна абсолютной разности годов рождения). Рассматриваемые метрики качества также инвариантны относительно выбранного начала отсчёта и направления (увеличение года рождения или увеличение возраста). Поэтому в рамках диссертационной работы атрибут год рождения отождествляется с атрибутом возраст. В работе этот атрибут обозначается как возраст.

Таблица 3 — Количественные характеристики наборов данных.

набор данных	**атрибут**	$\|A\|$	$\|Y\|$	$\|V\|$	$\|E\|$
twitter	род деятельности	9	4 625	53 199	1 167 488
	доход	55			
vk1	род деятельности	16	1 125	5 996	113 804
	пол	2	1 008		
	возраст	28	467		
vk2	пол	2	4 178	89 457	1 344 545
	возраст	45			
pokec	пол	2	1 138 314	1 632 803	22 301 964
	возраст	112			

2.3.1 Существующие наборы данных

Набор данных *twitter* [85] содержит социальный граф, собранный из социальной сети Twitter, и значения рода деятельности и дохода для некоторых пользователей. Для сбора данных в автоматическом режиме просматривались профили пользователей. Значения рода деятельности извлекались из поля «описание» профиля. Из описания профиля в свободной форме извлекались классы 1-го уровня из стандартного классификатора SOC[1] с использованием шаблонов и регулярных выражений. Доход каждого пользователей оценивался как средний доход для соответствующего класса 3-го уровня согласно классификатору SOC. Для полученного множества пользователей Y, размеченных родом деятельности и доходом, было собрано подмножество социального графа следующим образом. Для каждого из размеченных пользователей был получен список соседних страниц, или список страниц, на которые подписан данный пользователь. Затем все страницы, кроме входящих в исходное множество пользователей, которые связаны с менее чем $T = 10$ вершинами, были исключены из набора данных. В итоге граф $G = (V,E)$ содержит множество вершин V и рёбер $(x,z) \in E$, таких что хотя бы одна из инцидентных вершин является размеченной: $x \in Y$ или $z \in Y$. Формально данный процесс описан в алгоритме 3.

Набор данных *pokec* [84] является снимком социальной сети Pokec[2]. Он содержит все социальные связи между пользователями и профили из данной социальной сети, находящимися в открытом доступе. Профили содержат значения различных демографических атрибутов, включая пол и возраст. Данные значения были указаны самими пользователями социальной сети.

2.3.2 Набор данных со вручную размеченными значениями рода деятельности

Теперь опишем наборы данных, собранные автором данной работы. Сначала опишем набор данных, который содержит подмножество социального графа

[1] Standard Occupational Classification, классификация профессий, принятая в Великобритании

[2] https://pokec.azet.sk/

Вход : Y – множество размеченных вершин, T – порог фильтрации
Выход: $G(V,E)$ – подмножество социального графа

```
1  // формирование графа
2  V := {};
3  E := {};
4  for x ∈ Y do
5      V := V ∪ {x};
6      N_1 := getLinks(x); // получить множкство инцидентных вершин
7      for z ∈ N_1 do
8          V := V ∪ {z};
9          E := E ∪ {{x,z}}; // {x,z} - ненаправленное ребро
10     end
11 end
12 // фильтрация графа
13 for z ∈ V \ Y do
14     if |{{x,z} ∈ E : x ∈ Y}| < T then
15         E := E \ {{x,z} ∈ E : x ∈ Y};
16         V := V \ z;
17     end
18 end
19 return G = (V,E);
```

Алгоритм 3: Построение социального графа по множеству размеченных пользователей

социальной сети Вконтакте. Метки, представляющие значения рода деятельности, собирались вручную при помощи аннотаторов-добровольцев.

В качестве целевых аккаунтов для сбора данных рассматриваются только аккаунты студентов и выпускников МГУ. Это необходимо для уменьшения количества различных значений рода деятельности: рассматриваются только профессии, которые могут встретиться у выпускников вуза. Значения рода деятельности были взяты из Общероссийскиого классификатора видов экономической деятельности[3]. Было выбрано 19 наиболее возможных значений рода деятельности из данного классификатора, которые могут встретиться среди вы-

[3] http://www.consultant.ru/document/cons_doc_LAW_163320/

пускников МГУ и 20 наиболее популярных факультетов МГУ. Не смотря на то, что многие студенты старших курсов трудоустроены или проходят стажировку, мы предполагаем, что основным родом их деятельности является учёба. В качестве значения рода деятельности для студентов рассматривается уёеба на определённом факультете. Для выпускников в качестве значения рода деятельности рассматривается некоторый класс из Общероссийскиого классификатора видов экономической деятельности.

Набор данных собирался из социальной сети Вконтакте и формировался в два этапа. На первом этапе собиралось множество пользователей с размеченным родом деятельности. На втором этапе для полученного множества пользователей Y собирался социальный граф. Способ построения этого набора данных сбора аналогичен способу сбора набора данных *twitter*, отличие в том, что на первом этапе референсные значения рода деятельности собирались вручную, а при сборе *twitter* использовались автоматические методы.

Для сбора референсных значений рода деятельности было разработано веб-приложение. Аннотаторы использовали его, чтобы размечать значения рода деятельности и факультета для тех студентов и выпускников МГУ, которых они знают.

При входе на главную страницу аннотатор видит предложение авторизоваться в социальной сети Вконтакте и разрешить нашему приложению доступ к списку друзей. Доступ к этим данным позволяет приложению сформировать список кандидатов для разметки из друзей аннотатора, что упростит разметку для аннотатора: информацию о работе и учёбе своих друзей аннотатор скорей всего знает. Приложение отображает таблицу с аккаунтами друзей аннотатора, где для каждого друга указано значения факультета и рода деятельности. Пример таблицы изображен на рисунке 2.1. Изначально значения факультета и рода деятельности не заполнены, таблица заполняется в процессе разметки. Аннотатор может выбрать одного или несколько своих друзей из списка и нажать на кнопку «разметить». При этом отображается веб-страница, на которой аннотатору предлагается указать значения двух атрибутов: «факультет» и «род деятельности». Пример такой страницы изображен на рисунке 2.2. В качестве факультета предлагается выбрать тот, на котором учится, или который окончил выбранный пользователь или множество выбранных пользователей. Для каждого из атрибутов имеется возможность указать значение «без изменения». В этом случае если значение атрибута для пользователь было указано ранее,

Таблица

Открыть инструкцию

Пользователь	Выбрать	Действия	Факультет	Род деятельности	Статус
Андрей Гомзин	☐	Разметить X	Вычислительной математики и кибернетики	Научные исследования и разработки	Размечено
Денис Турдаков	☐	Разметить X	Вычислительной математики и кибернетики	Научные исследования и разработки	Размечено
Евгений Корныхин	☐	Разметить X	Вычислительной математики и кибернетики		Частично
Константин Архипенко	☐	Разметить X			

Рисунок 2.1 — Приложение для разметки рода деятельности. Пример страницы с таблицей

Выберите вуз и вид текущей деятельности

Открыть инструкцию

Пользователи: Андрей Гомзин Денис Турдаков

Факультет: Вычислительной математики и кибернетики

Деятельность: (M-72) Научные исследования и разработки

Сохранить

Информация по видам деятельности:

На сайте "Консультант Плюс"

При выборе "(Z-ZZ) Другое" укажите, пожайлуйста вид деятельности согласно данному классификатору.

Рисунок 2.2 — Приложение для разметки рода деятельности. Пример страницы выбора рода деятельности и факультета

оно не изменится. Опция «удалить» позволяет удалить ошибочное значение атрибута. После выбора значений атрибутов приложение показывает обновлённую таблицу. Аннотатор может продолжить размечать аккаунты. Результаты разметки сохраняются автоматически.

Один из возможных алгоритмов работы аннотатора с приложением заключается в следующем. Сначала аннотатор выбирает в таблице множество знакомых пользователей с одного факультета и один раз указывает факультет для всех выбранных аккаунтов. Процесс повторяется для других факультетов. Аналогично выбираются множества пользователей с одинаковым родом деятельности и указывается для всех них соответствующее значение.

В сборе референсных значений данного набора данных участвовало 69 аннотаторов. Так как один пользователь мог быть размечен несколькими аннотаторами, причём метки могли не совпадать, для получения наиболее правдоподобных значений атрибутов использовался простой итеративный алгоритм, схема которого описана в работе [102].

Каждый аннотатор i моделируется переменной $q_j \in [0, 1]$, означающей уровень доверия этому аннотатору, равный вероятности верного ответа. Предполагается, что все ответы аннотаторов независимы. Изначально уровни доверия аннотаров оцениваются одинаково. Затем выполняется несколько итераций. Метки для пользователей $Y = [y_1, \ldots, y_{|Y|}]$ вычисляются при заданных метках L, указанных аннотаторами, и фиксированными уровнями доверия аннотаторов $Q = [q_1, \ldots q_M]$. Затем уровни доверия аннотаторов Q оцениваются при наблюдаемых метках пользователей Y и метках L, указанных аннотатороами. Для оценки вероятностей используется сглаживание Лапласа с $\delta = 0.001$. Процесс повторяется до сходимости. В алгоритме 4 процесс описан формально.

Алгоритм оценки наиболее правдоподобных значений меток пользователей применялся независимо для каждого из исходных атрибутов «факультет» и «деятельность». После чего из значений этих атрибутов формировалось окончательное значение атрибута «род деятельности»: для студентов значением является факультет, для выпускников – деятельность. Наиболее редкие значения рода деятельности (не менее 10 пользователей с данной меткой) вместе с соответствующими пользователями удалялись из набора данных. Среднее значение коэффициента согласия Каппа Коэна между экспертными значениями и значениями, полученными с помощью алгоритма 4, около 0.86.

После сбора множества размеченных пользователей собирался социальный граф. Процесс сбора аналогичен процессу сбора графа из набора данных *twitter* и соответствует алгоритму 3. В качестве инцидентных вершин, возвращаемых функцией getLinks, рассматривались пользователи, с которыми дружит данный пользователь и публичные страницы групп, на которые подписан данный пользователь.

Обозначим полученный набор данных как *vk1*. Для того, чтобы использовать собранный социальный граф для оценки качества предсказания других атрибутов, дополнительно были получены значения возраста и пола. Для этой цели были собраны профили для размеченных пользователей. Из профилей были извлечены значения пола и возраста, в тех случаях, когда это возможно.

Вход : L – разреженная матрица ответов. $L_{ij} \in A \cup \{\varepsilon\}$ – ответ аннотатора $j \in [1, \ldots, M]$ для пользователя $i \in [1, \ldots, |Y|]$, $L_{ij} = \varepsilon$ – аннотатор j не размечал пользователя i.

Выход: Метки для пользователей $Y = [y_1, \ldots, y_{|Y|}]$ и уровень доверия аннотаторов $Q = [q_1, \ldots q_M]$

$Q := [0.8, \ldots, 0.8]$; // инициализация q_i некоторыми равными значениями

repeat

  for $i := 1$ **to** $|Y|$ **do**

    // наиболее правдоподобное y_i при заданных L и Q

    $y_i := \arg\max_{a \in A} \prod_{L_{ij}=a} q_j \prod_{L_{ij} \in A \setminus \{a\}} \frac{1-q_j}{|A|-1}$;

  end

  for $j := 1$ **to** M **do**

    // оценка q_j при заданных L и Y

    $q_j := \frac{\left|\{i : L_{ij} = y_i\}\right| + \delta}{\left|\{i : L_{ij} \neq \varepsilon\}\right| + |A|\delta}$;

  end

until *convergence*;

return Y, Q

Алгоритм 4: Итеративный алгоритм определения максимально правдоподобных значений меток для пользователей

2.3.3 Репрезентативный социальный граф со значениями атрибутов из профиля

Описанный выше набор данных *vk1* ограничен одним сообществом, представляющим студентов и выпускников МГУ, и имеет относительно небольшой размер. В связи с чем был собран еще один набор данных, представляющий более репрезентативное подмножество социальной сети Вконтакте.

Сначала собиралось множество пользователей, которые явно указали пол и возраст в своём профиле. Для этого использовался метод сэмплинга, основанный на модели Forest Fire [98]. В алгоритме 5 описан процесс сбора множества размеченных пользователей.

Вход : $G = (V, E)$ – социальный граф, N – ожидаемый размер сэмпла, p – вероятность поджога
Выход: S – целевое множество размеченных вершин

```
1  S := {};
2  B := {}; // множество подожженных вершин
3  Q := {}; // множество просмотренных вершин
4  while |S| < N do
5      if |B| = {} then
6          B := B ∪ { peek_random(V \ Q) } ;
7      end
8      // выбрать и обработать одну из подожженных вершин
9      x := peek_random(B) ;
10     B := B \ {x};
11     S := S ∪ {x}; // добавить вершину в результат
12     // обработать непросмотренных соседей вершины x
13     for n ∈ neighbors(G, x) \Q do
14         Q := Q ∪ {n};
15         if random([0, 1)) < p then
16             B := B \ {n}; // поджечь вершину
17         end
18     end
19 end
20 return S;
```

Алгоритм 5: Алгоритм Forest Fire для сбора множества размеченных пользователей

Для сбора набора данных в качестве значения параметра «вероятность поджога» использовалось значение 0.8. В нашем случае метод сэмплинга игнорирует неразмеченные узлы, «огонь» распространяется только по рёбрам между двумя размеченными вершинами. Алгоритм завершается, когда количество полученных размеченных пользователей достигнет 5000. Таким образом, количество размеченных пользователей будет близко к количеству размеченных пользователей в наборе данных *twitter*. С целью фильтрации аккаунтов с ложной информацией, из множества размеченных пользователей были исключены пользователи, родившиеся, как они указали, до 1960 года.

После сбора размеченных пользователей, собиралось подмножество социального графа аналогично наборам данных *twitter*, *vk1*. Неразмеченные вершины с наблюдаемым количеством рёбер менее $T = 10$ были исключены из набора данных. Обозначим полученный набор данных как *vk2*.

Собранные наборы данных *vk1* и *vk2* выложены в открытый доступ[4]. Идентификаторы вершин были захэшированы. Идентификаторы аннотаторов также были захэшированы.

2.4 Определение и исследование специфичности контекста

Можно выделить две идеи, на которых основаны существующие решения задачи предсказания значений демографического атрибута по социальному графу. Например, методы, основанные на распространении меток в графе, используют свойство «гомофилии». Оно заключается в том, социальные связи, представленные рёбрами социального графа, чаще создаются между пользователями с похожими демографическими характеристиками. Обозначим это свойство как *H*. Векторные представления вершин, используемые в других методах решения рассматриваемой задачи, основаны на идее похожести контекста. Векторные представления строятся таким образом, что вершины с более похожим контекстом имеют более близкие представления. Под контекстом вершины понимается множество вершин, с которыми она связана. Методы, основанные на подобных представлениях, предполагают, что вершины с более пересекаю-

[4] https://nextcloud.ispras.ru/s/derYJWQT2seaZeg

щимся контекстом имеют более похожие значения атрибута. Обозначим данное свойство как C.

В рамках диссертационной работы предлагается новое свойство, *специфичность контекста*, обозначаемое как CS. Идея данного свойства заключается в следующем. Если окрестность вершины состоит преимущественно из пользователей с одинаковым значением атрибута, предполагается, что данная вершина более полезна для предсказания значений данного атрибута для пользователей. Наоборот, если окрестность вершины представляет собой генеральную совокупность по значениям данного атрибута, данная вершина не играет существенную роль в решении данной задачи. Рассмотрим, например, публичную страницу, посвященную некоторому яхт-клубу. Если среди подписчиков, у которых известен доход, преимущественно встречаются пользователи с высоким доходом, то можно предположить, что остальные пользователи также имеют высокий доход. Другими словами, подписка на данное сообщество – важный признак для предсказания уровня дохода пользователей. Рассмотрим другую публичной страницу, у которой значения рассматриваемого атрибута среди подписчиков близко к генеральной совокупности, например, некоторый новостной ресурс. Подписка на такое сообщество имеет меньшую специфичность и даёт мало информации для предсказания значений атрибута пользователей.

Сначала определяется понятие специфичности контекста для вершины и специфичности общего контекста для пары вершин. Затем исследуются «гомофилия» (H) и зависимости между свойствами общего контекста (C и CS) и значениями атрибута в наборах данных. Под «гомофилией» (H) понимается зависимость между наличием или отсутствием ребра и значениями атрибута вершин. В качестве свойств общего контекста рассматриваются размер (C) и специфичность (CS) общего контекста.

2.4.1 Специфичность контекста для вершины и общего контекста для пары вершин

Определим значение специфичности контекста $s(z)$ для заданной вершиныz следующим образом:

$$s(z) = \begin{cases} KL\big(d(N_z)\big|\big|d(Y)\big) & \text{, если } \exists d(N_z) \\ 0 & \text{, иначе} \end{cases} \tag{2.2}$$

Здесь $KL(p||q) = \sum_i \log p_i \frac{p_i}{q_i}$ – дивергенция Кульбака–Лейблера.

Специфичность контекста $s(z)$ показывает, насколько распределение $d(N_z)$ значений $\mathcal{A}$ для соседей вершины z отличается от распределения генеральной совокупности $d(Y)$, т.е. распределения для всех размеченных вершин Y. Если распределение меток соседей близко к распределению генеральной совокупности, то специфичность вершины близка к 0. Чем больше расстояние между этими распределениями, тем больше значение специфичности вершины.

Для заданной пары вершин (x, z) определим специфичность как:

$$cs(x, z) = \sum_{v \in N_x \cap N_z} s(v) \tag{2.3}$$

Специфичность пары вершин (x, z) является численной характеристикой общего контекста данных вершин, то есть множества вершин, связанных как с x, так и с y. Данная характеристика зависит как от количества общих соседей, так и от специфичности контекста $s(v)$ каждой вершины их общих соседей $N_x \cap N_z$. Чем больше общих соседей с большим значением $s(v)$, тем больше специфичность контекста для пары вершин $cs(x, z)$. Данную величину также будем называть **специфичностью общего контекста** вершин.

2.4.2 Исследование «гомофилии» и зависимостей между свойствами общего контекста и значениями атрибута в наборах данных

Введём несколько дополнительных обозначений и формализуем свойства H, C и CS и их значения.

Напомним, что набор данных включает в себя подмножество социального графа $G = (V,E)$, множество размеченных вершин $Y \subseteq V$ и метки (значения атрибута) $y_x : x \in Y$. Обозначим множество всех возможных пар размеченных вершин $\Omega = \{(x,z) : x,z \in Y\}$.

Рассмотрим вероятностное пространство с множеством элементарных исходов Ω, сигма-алгеброй $\sigma = 2^{\Omega}$ и функций вероятности $\mathbb{P}(\omega) = \frac{1}{|\Omega|}, \forall \omega \in \Omega$.

Для каждой пары вершин (x,z) введем функции $h(x,z)$, $c(x,z)$ и $cs(x,z)$, определяющие значения свойств H, C и CS, соответственно. Значение свойства H определим как *наличие ребра* между вершинами:

$$h(x,z) = \begin{cases} 1 & \text{, если } (x,z) \in E \\ 0 & \text{, иначе} \end{cases} \tag{2.4}$$

Значение свойства C определим как *размер общего контекста* вершин:

$$c(x,z) = |N_x \cap N_z| \tag{2.5}$$

Значение свойства CS определим как *специфичность общего контекста*, это значение описано формулой (2.3).

Значения h, c и cs вычисляются на наборах заданных, описанных в разделе 2.3 и анализируются. Целями анализа которого являются:

- проверить, увеличивается ли вероятность совпадения значений атрибута двух вершин с увеличением значений h, c и cs;
- сравнить, насколько сильно данное свойство выражено для каждой из величин h, c и cs;
- оценить, для какой части данных это наблюдается.

Опишем процесс анализа формально. Для заданного набора данных для каждой пары размеченных вершин $(x,z) \in \Omega$ вычисляются значения $h(x,z)$, $c(x,z)$, $cs(x,z)$. При вычислении cs дробные значения округляются вниз с точностью до 0.001. Для каждой из случайных величин h, c и cs упорядочим по возрастанию множества их значений. Получим три последовательности значений: $(h_i)_{i=1}^{n_h}$, $(c_i)_{i=1}^{n_c}$, $(cs_i)_{i=1}^{n_{cs}}$. Здесь n_h, n_c и n_{cs} – количество различных значений величин h, c и cs, соответственно, полученных и данных.

Пусть ξ – одна из случайных величин h, c или cs. Для каждого из значений, которое может принимать случайная величина ξ, рассмотрим множество вершин $\{(x,z) \in \Omega : \xi(x,z) \geqslant \xi_i\}$ и определим для него:

$$\alpha_i^{\xi} = P(\xi(x,z) \geqslant \xi_i), \text{ где } i = \overline{1,n_\xi} \tag{2.6}$$

Множество пар вершин, соответствующих значению α_i^{ξ}, обозначим как $B_\xi(\alpha_i^{\xi}) = \{(x,z) \in \Omega : \xi(x,z) \geqslant \xi_i\}$. Легко видеть, что $\alpha_1^{\xi} = 1$, с увеличением ξ_i значение α_i^{ξ} уменьшается. Значения α_i^{ξ} показывают, в какой пропорции заданное значение h_i, c_i или cs_i делит множество пар размеченных вершин Ω на пары $B_\xi(\alpha_i^{\xi})$ с равным или более высокими и пары $\Omega \setminus B_\xi(\alpha_i^{\xi})$ с более низкими значениями соответствующей случайной величины.

Теорема 1 Пусть диаметр графа $D(G) > 2, \exists\, v \in V : |N_v| > 1, s(v) > 0$. Тогда $h_1 = c_1 = cs_1 = 0$ и $\exists h_2 > 0, c_2 > 0, cs_2 > 0$; при этом α_2^h является плотностью графа; α_2^c равно вероятности того, что для случайной пары вершин существует путь длины 2 между ними.

Доказательство. Так как $D(G) > 1$, то $\exists x,z \in V, x \neq z : (x,z) \notin E$. Для таких пар вершин $h = 0$, следовательно, $h_1 = 0$.

Из $D(G) > 2$ следует, что $\exists x, z \in V, x \neq z : d(x,z) > 2$, где $d(x,z)$ – длина кратчайшего пути между вершинами. Для таких вершин $c = 0$ и $cs = 0$, следовательно, $c_1 = cs_1 = 0$.

Из условия $\exists\, v \in V : |N_v| > 0$ следует, что $|E| > 0$. Отсюда получаем, что для графа существует $h_2 = 1$. По определению (2.6): $\alpha_2^h = P\big(h(x,z) \geqslant 1\big) = P\big(h(x,z) = 1\big)$. Число элементарных исходов равно $\frac{|V|\big(|V|-1\big)}{2}$, число исходов, удовлетворяющих условию $h(x,y) = 1$ равно $|E|$. Таким образом, получим: $\alpha_2^h = \frac{2|E|}{|V|\big(|V|-1\big)}$, что является плотностью графа.

Из условия $\exists\, v \in V : |N_v| > 1$ следует, что $\exists x,z \in V, x \neq z : v \in N_x \cap N_z$, то есть $c(x,z) = |N_x \cap N_z| > 0$. Отсюда следует, что существует значение $c_2 > 0$. С учётом $s(v) > 0$, так как v является частью общего контекста x и z, то $cs(x,z) \geqslant s(v) > 0$, следовательно, существует значение $cs_2 > 0$. Из (2.6) и $c_1 = 0$ получим: $\alpha_2^c = P(c(x,z) \geqslant c_2) = P\big(c(x,z) > 0\big)$, что является вероятностью того, для случайной пары вершин (x,z) выполняется $|N_x \cap N_z| > 0$, то есть существует путь длины 2 между ними. □

Для всех рассматриваемых наборов данных выполняются условия теоремы 1: $D(G) > 2$, $\exists v \in V : |N_v| > 1, s(v) > 0$. Следовательно, для каждой из величин h, c, cs существует не менее двух различных значений. Таким образом, можно проводить анализ наборов данных на предмет связи этих величин со значениями атрибута вершин. Дальнейшие рассуждения предполагают выполнение условий теоремы 1.

Для значений ξ_i определим:

$$r_i^{\xi} = r(\alpha_i^{\xi}) = P(y_x = y_z | \xi(x,y) \geqslant \xi_i), \text{ где } i = \overline{1,n_{\xi}} \tag{2.7}$$

r_i^{ξ} показывают долю пар вершин с совпадающим значением метки среди заданного множества пар с большими значениями h, c, cs, в зависимости от рассматриваемого свойства. Стоит отметить, что значения возраста и дохода считаются равными, если они точно совпадают. При анализе данных, описываемом в данном подразделе похожесть и разница неравных значений атрибутов, предсказание которых сводится к задаче регрессии (возраст и доход), не учитывается. Количество различных значений атрибутов см. в таблице 3.

Далее опишем, каким образом вычисляются α_i^{ξ} и r_i^{ξ}. Для заданного значения случайно величины $\xi = \xi_i$ рассмотрим множество пар с соответствующим значением ξ. Пусть $[=]_i^{\xi}$ – количество пар из этого множества, у которых которых совпадает значение атрибута, т.е. $y_x = y_z$; $[\neq]_i^{\xi}$ – количество пар из этого множества, для которых $y_x \neq y_z$. Соответствующие значения α_i^{ξ} и r_i^{ξ} вычисляются следующим образом:

$$\alpha_i^{\xi} = \frac{\sum\limits_{j=i}^{n_{\xi}} [=]_j^{\xi} + \sum\limits_{j=i}^{n_{\xi}} [\neq]_j^{\xi}}{\sum\limits_{j=1}^{n_{\xi}} [=]_j^{\xi} + \sum\limits_{j=1}^{n_{\xi}} [\neq]_j^{\xi}} \tag{2.8}$$

$$r_i^{\xi} = r(\alpha_i^{\xi}) = \frac{\sum\limits_{j=i}^{n_{\xi}} [=]_j^{\xi}}{\sum\limits_{j=i}^{n_{\xi}} [=]_j^{\xi} + \sum\limits_{j=i}^{n_{\xi}} [\neq]_j^{\xi}} \tag{2.9}$$

Таким образом, для каждой из величин h, c, cs и для каждого её значения (h_i, c_i cs_i соответственно) ставится в соответствие два значения: α_i – доля пар со значением соответствующего свойства не менее ξ_i и r_i – вероятность совпадения меток среди тех пар, у которых значение соответствующего свойства не менее ξ_i.

Определим также референсное значение $R = P(y_x = y_z)$. Данное значение есть вероятность того, что две случайно выбранные размеченные вершины имеют одинаковую метку (значение атрибута). Пусть $\#[a_i]$ – количество вершин с меткой a_i. Тогда R вычисляется следующим образом:

$$R = \sum_{i=1}^{|A|} \frac{\#[a_i]\big(\#[a_i] - 1\big)}{2} \tag{2.10}$$

Утверждение 1 $R = r_1^h = r_1^c = r_1^{cs}$.

Доказательство. Следует из определения $r_1^\xi = P(y_x = y_z | \xi \geqslant \xi_1)$. Так как ξ_1 – минимальное значение, которое может принимать ξ, то $r_1^\xi = P(y_x = y_z)$. □

Построим график, показывающий изменение вероятности совпадения значений атрибута при увеличении значений величин h, c, cs. С увеличением этих значений α^ξ будет уменьшаться. Поэтому на оси абсцисс расположим значения $1-\alpha_\xi$, по оси ординат расположим значения r^ξ. Изобразим точку (α_2^h, r_2^h) на графике. Аналогично изобразим точки (α_i^c, r_i^c), $i \in \overline{2, n_c}$ и точки $(\alpha_i^{cs}, r_i^{cs})$, $i \in \overline{2, n_{cs}}$. Из теоремы 1 и утверждения 1 следует, что $(\alpha_1^h, r_1^h) = (\alpha_1^h, r_1^h) = (\alpha_1^h, r_1^h) = (1, R)$. Для большей наглядности вместо трёх одинаковых точек отобразим референсное значение R в виде горизонтальной пунктирной линии. Значение для h представлено чёрным квадратом, значения для c – синими треугольниками, значения для cs – зелеными кругами. Графики для различных наборов данных и атрибутов представлены на рисунках 2.3-2.11.

Графики позволяют наглядно показать зависимость между значениями вероятности совпадения значений атрибута при различных разбиениях множества пар по значениям h, c и cs. Однако графики не показывают статистическую значимость этих зависимостей. Для оценки статистической значимости проведём дополнительный эксперимент.

Значению α_i^ξ однозначно соответствует множество пар $B_\xi(\alpha_i^\xi) \subset \Omega$. Определим способ получения множества пар вершин $B_\xi(\alpha)$, соответствующей произвольному значению $\alpha \in (\alpha_{i+1}^\xi, \alpha_i^\xi)$. Для этого воспользуемся следующей интерполяцией. Рассмотрим множество $B_\xi(\alpha_{i+1}^\xi)$ и дополним его случайными парами из множества $\{(x,z) \in \Omega : \xi(x,z) = \xi_i\}$. Количество дополнительных пар равно $\lfloor |\Omega|(\alpha - \alpha_{i+1}^\xi) \rfloor$.

Значения $r^\xi = r^\xi(\alpha) = P\Big(y_x = y_z | B_\xi(\alpha)\Big)$ для произвольного значения $\alpha \in (\alpha_{i+1}^\xi, \alpha_i^\xi)$ вычисляются следующим образом:

$$r^\xi(\alpha) = \frac{\frac{\alpha - \alpha_{i+1}^\xi}{\alpha_i^\xi - \alpha_{i+1}^\xi}[=]_i^\xi + \sum\limits_{j=i+1}^{n_\xi}[=]_j^\xi}{\frac{\alpha - \alpha_{i+1}^\xi}{\alpha_i^\xi - \alpha_{i+1}^\xi}[=]_i^{\xi_i} + \frac{\alpha - \alpha_{i+1}^\xi}{\alpha_i^\xi - \alpha_{i+1}^\xi}[\neq]_i^\xi + \sum\limits_{j=i+1}^{n_\xi}[=]_j^\xi + \sum\limits_{j=i+1}^{n_\xi}[\neq]_j^\xi} \tag{2.11}$$

Описанная интерполяция предполагает равномерное (по совпадению значений меток) упорядочивание пар вершин с одинаковым значением ξ. Дробь

$\frac{\alpha-\alpha^{\xi}_{i+1}}{\alpha^{\xi}_{i}-\alpha^{\xi}_{i+1}}$ показывает долю пар с $\xi = \xi_i$, участвующих в оценке значения r: при α близком к α_{i+1} эта доля минимальна, при α близком к α^{ξ}_{i} – максимальна.

Полученные выборки использовались для оценки связи между значением свойств h, c, cs и совпадением значения атрибута для пар вершин. Кроме того, сравнивалось, насколько частота совпадения значений атрибута среди вершин с наибольшими значениями свойства cs выше, чем частота среди вершин с наибольшими значениями h и c, при одинаковом размере выборок. Для этой цели были проведены статистические t-тесты Стьюдента для различных выборок при некоторых фиксированных значениях $(1 - \alpha) \in \{0.25, 0.5, 0.75, 0.8, 0.95, 0.99, 0.999\}$. Сравнивались выборки $B_\xi(\alpha)$ и Ω для каждой из случайных величин h, c, cs. Также сравнивались выборки $B_{cs}(\alpha)$ и $B_h(\alpha)$, $B_{cs}(\alpha)$ и $B_c(\alpha)$. В рамках одного теста для двух рассматриваемых выборок формулировалась нулевая гипотеза H_0, заключающаяся в том, что совпадение значений атрибутов у пары вершин (x,z) не зависит от выборки, затем для пар из сравниваемых выборок применялся t-тест Стьюдента к значениям случайной величины, равной 1, если $y_x = y_z$, и равной 0 в противном случае.

В таблицах 4-12 представлены значения $r\xi$ для характеристик h, c и cs для фиксированных значений $(1-\alpha) \in \{0.25, 0.5, 0.75, 0.8, 0.95, 0.99, 0.999\}$, а также значения p-value, полученные при применении t-тестов Стьюдента. Неравномерность выбранных значений $(1 - \alpha)$ связано со степенным распределением значений ξ в графах социальных сетей. Результаты теста показали, что выборки $B_h(\alpha)$, $B_c(\alpha)$ и $B_{cs}(\alpha)$ статистически значимо (с p-value менее 1%) отличаются от всех пар вершин Ω по совпадению значений атрибута, на всех рассматриваемых наборах данных. Во всех случаях, кроме свойства h в наборе данных pokec для атрибута пол, наблюдается, что значения атрибута для пар из B_ξ совпадают чаще, чем для произвольной пары из Ω. Также в большинстве случаев тесты показывают, что совпадение значений атрибута на выборках $B_{cs}(\alpha)$ происходит значимо чаще, чем на выборках $B_h(\alpha)$ и $B_c(\alpha)$.

Из графиков на рисунках 2.3-2.11 и значений из таблиц 4-12 можно сделать следующие выводы:

– На всех графиках для c и cs наблюдается преимущественный рост r с ростом α. Следовательно, вероятность совпадения значений атрибутов

случайной пары вершин растет при увеличении размера общего контекста и специфичности общего контекста.

- На всех графиках зеленые круги расположены выше синих треугольников при близких значениях α. Это означает, что для фиксированной доли пар с большим значением cs вероятность совпадения значений атрибутов больше, чем аналогичный показатель для c.
- В единичных случаях (атрибуты occupation и age в наборе данных vk1) значения h_2 на графиках высоки и сопоставимы с соответствующими значениями c и cs. Это свидетельствует о выраженной «гомофилии» для данных атрибутов и наборов данных.
- Во всех таблицах в большинстве случаев значения CS не хуже, чем соответствующие значения C.
- Для набора данных vk1 и атрибутов age и occupation наблюдается, что значения H превосходят C и CS при $1 - \alpha > 0.99$. Это свидетельствует о наличии «гомофилии» в наборе данных для обозначенных атрибутов.
- Описанные выше выводы основаны на статистически значимых результатах тестов.

Результаты данного анализа позволяют сделать вывод о целесообразности использования данных свойств в методах предсказании значений демографических атрибутов пользователей. Специфичность общего контекста показывает наибольшую корреляцию с совпадением значений атрибута пользователей, по сравнению с наличием ребра или размером общего контекста.

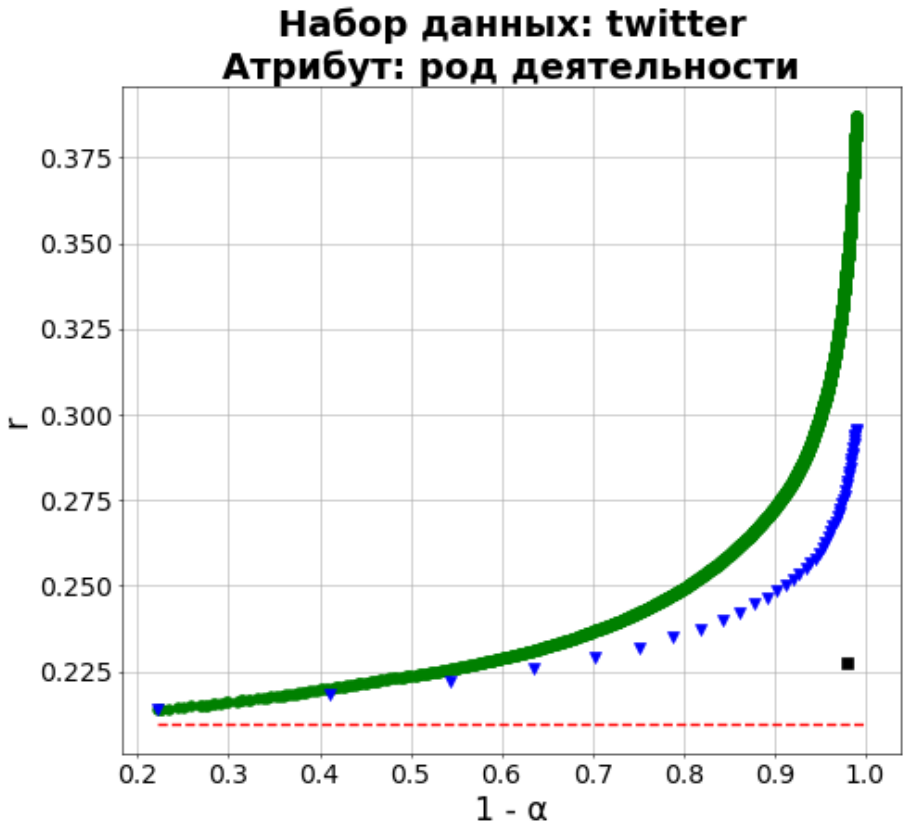

Рисунок 2.3 — Результаты анализа набора данных twitter; атрибут: род деятельности

Таблица 4 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .20988$; набор данных: twitter; атрибут: род деятельности

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.21000	.21023	.21094	.21306	.21661	.22772	.22772
p-value (r^h vs R)	.53494	.10070	.00014	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00001
значение r^c	.21433	.22062	.23199	.24804	.26003	.29632	.35259
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.21456	.22377	.24212	.27267	.29915	.38691	.55588
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	.25749	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

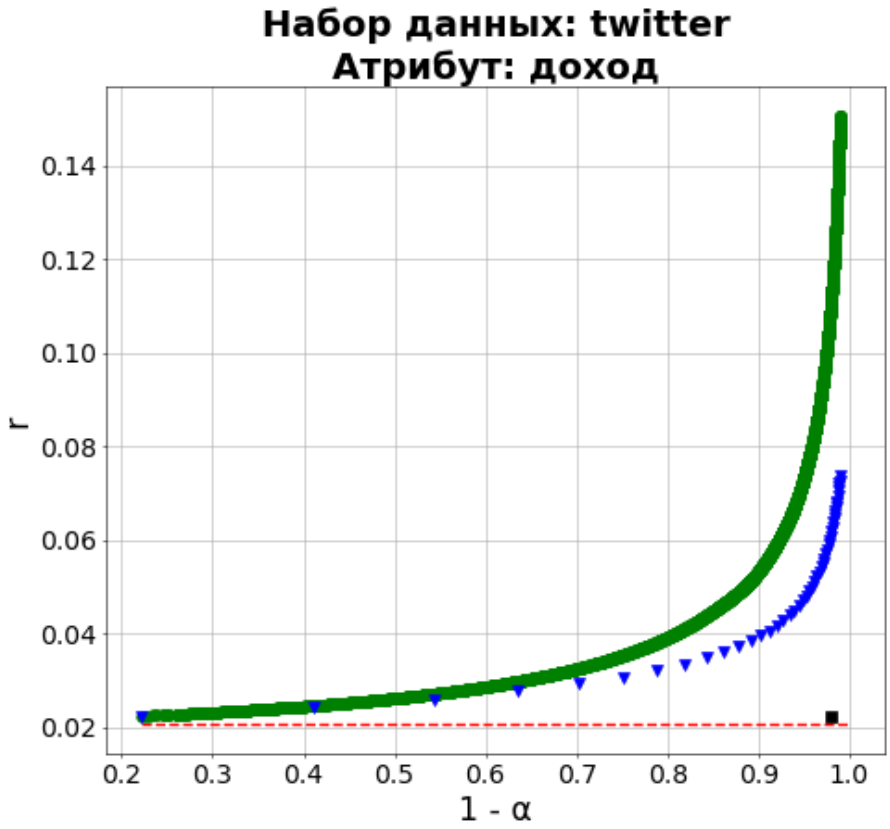

Рисунок 2.4 — Результаты анализа набора данных twitter; атрибут: доход

Таблица 5 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .02077$; набор данных: twitter; атрибут: доход

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.02078	.02079	.02085	.02103	.02132	.02222	.02222
p-value (r^h vs R)	.88483	.70159	.37333	.07180	.00595	.00088	.29037
значение r^c	.02253	.02523	.03063	.03921	.04733	.07396	.11741
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.02258	.02610	.03521	.05314	.07326	.15045	.31656
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	.49056	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

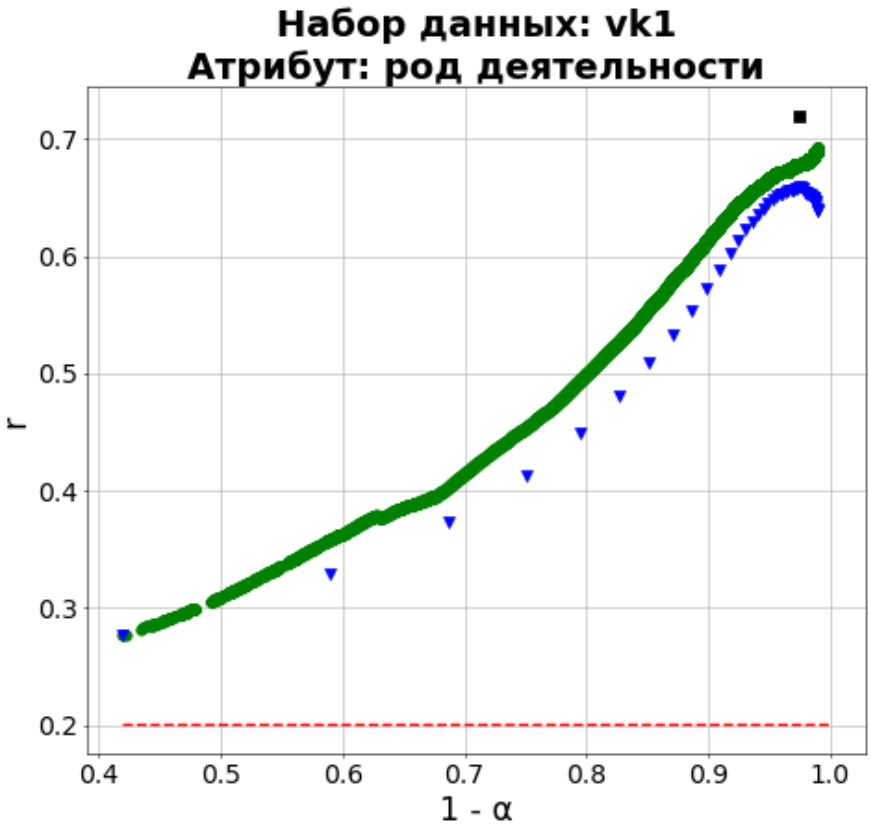

Рисунок 2.5 — Результаты анализа набора данных vk1; атрибут: род деятельности

Таблица 6 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .20141$; набор данных: vk1; атрибут: род деятельности

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.20607	.21537	.24328	.32702	.46659	.71899	.71899
p-value (r^h vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^c	.23599	.29681	.41281	.57494	.64588	.63812	.63769
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.23599	.30916	.45336	.61546	.66586	.69055	.71688
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00045	.93361
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00255

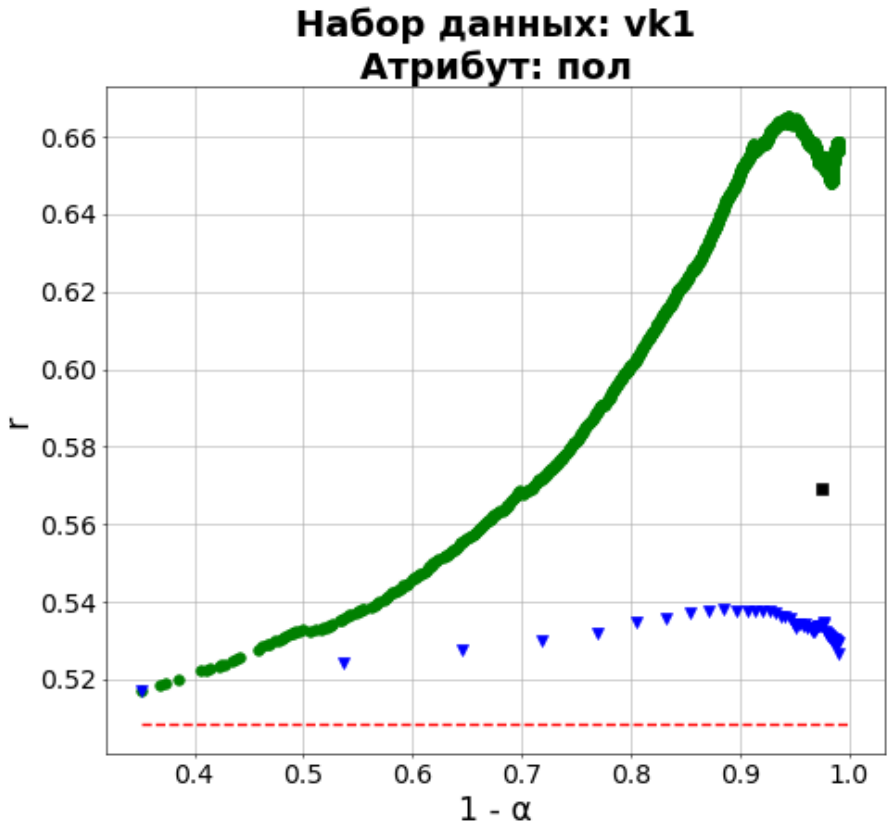

Рисунок 2.6 — Результаты анализа набора данных vk1; атрибут: пол

Таблица 7 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .50861$; набор данных: vk1; атрибут: пол

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.50914	.51021	.51339	.52294	.53885	.56941	.56941
p-value (r^h vs R)	.62073	.19047	.00235	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00618
значение r^c	.51389	.52236	.53122	.53752	.53403	.52704	.57810
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00897	.00175
значение r^{cs}	.51389	.53278	.58162	.65087	.66392	.65885	.72021
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	.00003	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

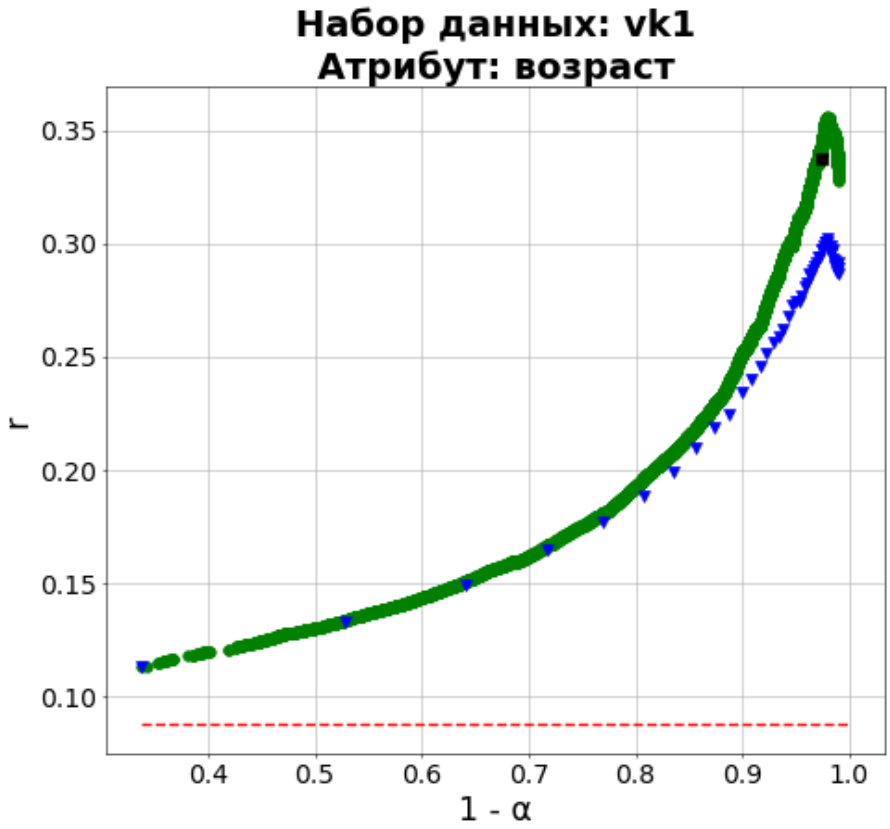

Рисунок 2.7 — Результаты анализа набора данных vk1; атрибут: возраст

Таблица 8 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .08825$; набор данных: vk1; атрибут: возраст

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.09042	.09476	.10777	.14680	.21186	.33767	.33767
p-value (r^h vs R)	.10037	.00002	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^c	.10437	.12922	.17194	.23458	.27456	.28997	.24081
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.10437	.13032	.17558	.25181	.30732	.32809	.28316
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.63541	.38497
p-value (r^{cs} vs r^c)	$< 10^{-5}$	.59093	.26321	.00304	.00017	.05422	.47770

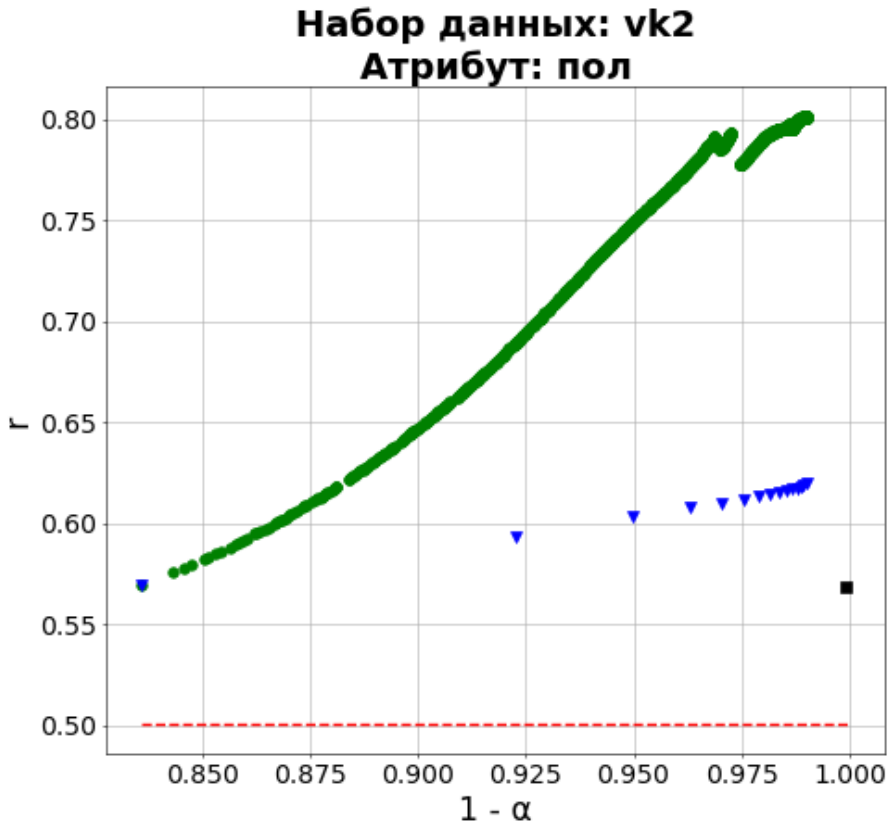

Рисунок 2.8 — Результаты анализа набора данных vk2; атрибут: пол

Таблица 9 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .50081$; набор данных: vk2; атрибут: пол

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.50083	.50087	.50099	.50134	.50193	.50666	.55986
p-value (r^h vs R)	.93925	.84021	.63940	.34330	.14763	.00058	$< 10^{-5}$
значение r^c	.50529	.51425	.54113	.58305	.60311	.61919	.65877
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.50529	.51425	.54113	.64684	.74897	.80091	.84758
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

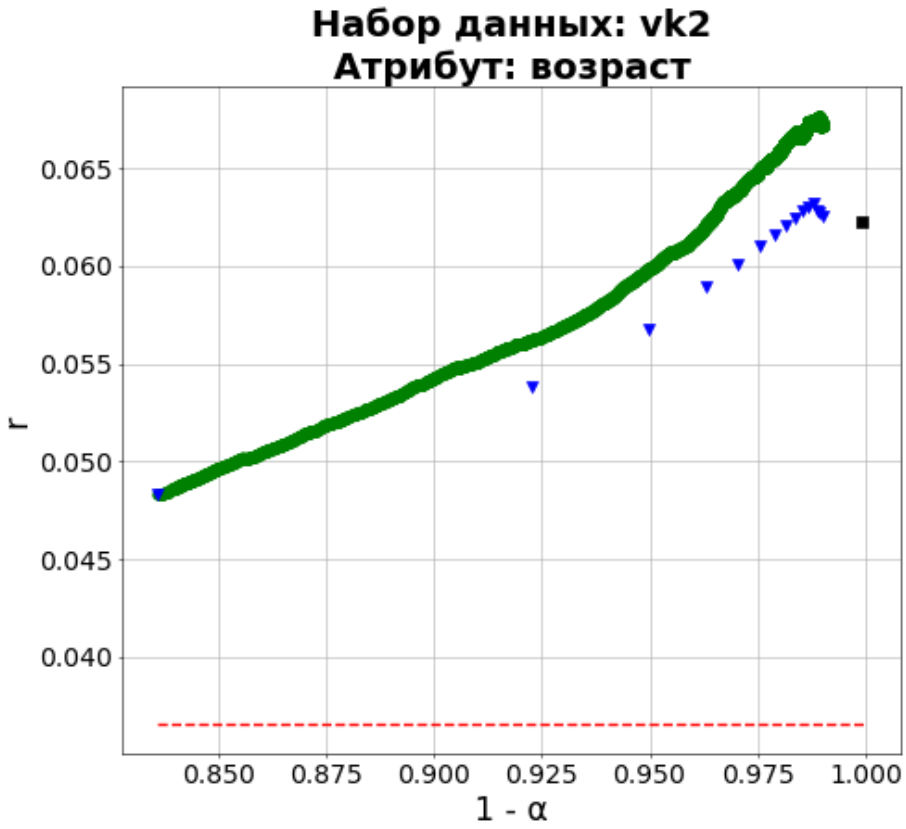

Рисунок 2.9 — Результаты анализа набора данных vk2; атрибут: возраст

Таблица 10 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .03661$; набор данных: vk2; атрибут: возраст

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.03662	.03664	.03668	.03682	.03704	.03883	.05900
p-value (r^h vs R)	.93871	.83880	.63643	.33906	.14410	.00052	$< 10^{-5}$
значение r^c	.03738	.03891	.04349	.05149	.05679	.06261	.05815
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.03738	.03891	.04349	.05427	.05987	.06723	.06085
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.60518
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	.00009	.45034

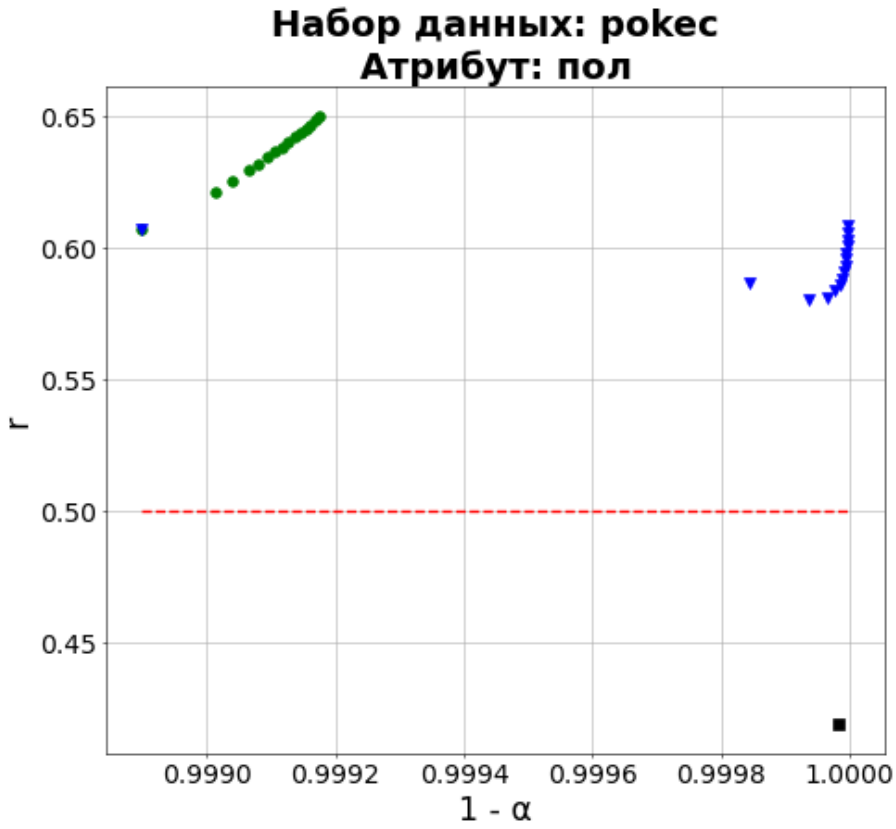

Рисунок 2.10 — Результаты анализа набора данных pokec; атрибут: пол

Таблица 11 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .50009$; набор данных: pokec; атрибут: пол; наблюдается, что $h_2 < R$

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.50009	.50009	.50009	.50008	.50007	.49996	.49875
p-value (r^h vs R)	.63775	.21285	.00379	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^c	.50013	.50021	.50045	.50116	.50234	.51179	.60685
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.50013	.50021	.50045	.50116	.50234	.51179	.61965
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

Рисунок 2.11 — Результаты анализа набора данных pokec; атрибут: возраст

Таблица 12 — Значения r для свойств h, c, cs при различных $1-\alpha$; $R = .03605$; набор данных: pokec; атрибут: возраст

$1-\alpha$	25.0%	50.0%	75.0%	90.0%	95.0%	99.0%	99.9%
значение r^h	.03605	.03605	.03606	.03607	.03610	.03630	.03857
p-value (r^h vs R)	.01762	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^c	.03608	.03613	.03628	.03675	.03753	.04378	.10750
p-value (r^c vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
значение r^{cs}	.03608	.03613	.03628	.03675	.03753	.04378	.11185
p-value (r^{cs} vs R)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^h)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
p-value (r^{cs} vs r^c)	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

2.5 Описание подхода для предсказания значений демографических атрибутов

Подход заключается в вычислении специфичности контекста для каждой из вершин социального графа и использования полученных значений в качестве веса, количественной характеристики важности связи с этой вершины, представленной ребром.

Способ использования этой характеристики задаётся конкретным методом. В следующем разделе в рамках подхода представлены методы предсказания значений атрибута для пользователей, основанные на специфичности контекста.

2.6 Выводы

В главе сформулировано и определено свойство специфичности контекста для вершин размеченного социального графа. Специфичность контекста показывает, насколько значения заданного атрибута среди соседей заданной вершины выделяются относительно распределения генеральной совокупности.

Были введены и формализованы значения, позволяющие оценить гомофилию, зависимость между *размером* общего контекста вершин и значениями атрибутов, зависимость между *специфичностью* общего контекста и значениями атрибутов пар вершин. Были описаны некоторые свойства этих значений и условия, при которых они выполняются. Эти свойства и условия сформулированы в виде теоремы, приведено доказательство теоремы.

На нескольких наборах данных, собранных из реальных социальных сетей, экспериментально показано, что специфичность контекста может быть использована для предсказания значений атрибутов. Кроме того, зависимость между специфичностью общего контекста и значениями атрибутов пар вершин на рассматриваемых наборах данных более выражена, чем гомофилия и зависимость между размером общего контекста вершин и значениями атрибутов.

Предложен подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа.

Собраны, анонимизированы и выложены в открытый доступ два набора данных из социальной сети Вконтакте. Один из набор содержит вершины, соответствующие студентам и выпускникам МГУ им. М.В. Ломоносова, социальные связи и значения рода деятельности, собранные вручную при помощи аннотаторов. Дополнительно набор данных содержит указанные в публичном профиле значения пола и возраста. Второй набор представляет репрезентативную выборку пользователей с заполненными атрибутами пол и возраст публичного профиля.

Глава 3. Методы предсказания значений атрибутов пользователей с использованием специфичности контекста

В главе описываются методы предсказания значений атрибутов пользователей. Методы разработаны в рамках подхода к предсказанию значений атрибутов на основе специфичности контекста. Проводится оценка вычислительной сложности методов, обсуждаются их особенности, проводится экспериментальное сравнение разработанных методов с существующими. В конце главы приводятся рекомендации к использованию разработанных методов, полученные на основе теоретической и экспериментальной оценки методов.

3.1 Методы на основе специфичности контекста

Представим несколько методов на основе специфичности контекста. Два метода, *LP-CS* и *LP-CS-Gen*, являются модификациями алгоритма распространения меток. Для метода *Distr2-CS-XGB* вводятся новые признаки для вершин *Distr2-CS*, являющиеся взвешенными распределениями значений $\mathcal{A}$ среди вершин из двухшаговой окрестности. В основе метода *GConv-CS[n]* лежит свёрточная графовая нейронная сеть.

3.1.1 LP-CS: модификация алгоритма распространения меток

LP-CS представляет собой модифицированный 2-шаговый синхронный алгоритм распространения меток. В алгоритме 2 описана схема работы метода. Метод имеет две итерации, для каждой из итераций опишем функцию select(), задающую способ выбора меток на каждом из шагов.

На первом шаге алгоритма для каждой вершины $z \in V$ вычисляются специфичность $s(z)$ и метка $v(z)$. Специфичность вычисляется согласно формуле (2.2). Метка $v(z)$ для вершины z выбирается как в базовом алгоритме:

самая частая метка среди соседей в случае задачи классификации, среднее значение в случае регрессии.

На втором шаге предсказываемое значение $p(x)$ атрибута $\mathcal{A}$ для вершины x вычисляется как метка, соответствующая максимальной сумме значений специфичности среди соседних вершин вершины x:

$$p(x) = \arg\max_{a_i \in A} \sum_{z \in N_x} \mathbb{1}_{v(z)=a_i} \cdot s(z) \tag{3.1}$$

Значения $p(x)$ для вершин $x \in V$ являются выходом *LP-CS* алгоритма.

Интуиция данного алгоритма заключается в следующем. На первом шаге известные значения атрибута y_x распространяются к вершине z от её соседей $x \in N_z$ и сохраняются в виде характеристики контекста вершины z – наиболее специфичным (выделяющимся) значением атрибута соседей $v(z)$ и количественной мерой данной специфичности $s(z)$. Так как каждая вершина x является частью контекста своих соседей, предсказываемое для неё значение $p(x)$ оценивается на втором шаге алгоритма путём «обратного распространения» вычисленных на первом шаге характеристик контекста $v(z)$ и $s(z)$. Значение атрибута $p(x)$ выбирается как метка с наибольшим весом среди меток контекста соседних вершин N_x: каждое значение $v(z)$ учитывается с весом, равным $s(z)$. Таким образом, алгоритм считает подписку на публичную страницу, посвящённую, например, военной тематике, более значимой, чем подписку на страницу артиста, популярного и среди мужчин, и среди женщин, для предсказания неизвестных значений пола пользователей.

3.1.2 LP-CS-Gen: алгоритм распространения меток, устойчивый к неравномерному распределению значений атрибута

Описанный выше метод *LP-CS* обладает недостатком, унаследованным от базовой версии. При применении алгоритма для решения задач классификации выбирается самая частая метка среди меток соседей. Данное решение плохо работает в случае, когда генеральная совокупность значений метки распределена неравномерно. Рассмотрим пример набора данных, в котором для 90% размеченных вершин значение метки равно a, а для 10% равно b. Допустим, что у

некоторой вершины 30 соседей с меткой a и 20 соседей с меткой b. *LP-CS* выберет метку a, так как она самая частая. Однако стоит отметить, что метка b встречается у соседей чаще, чем в генеральной совокупности, метка a, поэтому, имеет смысл выбрать метку b в качестве метки, характеризующей окружение вершины. Опишем модификацию алгоритма *LP-CS*, которая выбирает метку вершины с учетом генеральной совокупности.

На первом шаге алгоритма метка $v(z)$ для вершины z выбирается таким образом, что распределение значений атрибута среди соседей $d(N_z)$ максимально превышает распределение генеральной совокупности $d(Y)$ в выбранной точке:

$$v(z) = \arg\max_{a_i \in A} kl_i\big(d(N_z)\big|d(Y)\big) \tag{3.2}$$

Здесь $kl_i = \log p_i \frac{p_i}{q_i}$ – член суммы в формуле дивергенции Кульбака–Лейблера в случае дискретных распределений, соответствующий значению атрибута a_i. Данный метод назовём *LP-CS-Gen*.

3.1.3 Distr2-CS-XGB: метод на основе распределений значений атрибута на двухшаговой окрестности

Введём новые признаки *Distr2-CS* для представления вершин размеченного графа и метод для предсказания значений атрибута $\mathcal{A}$, основанный на этих признаках.

Признаковый вектор *Distr2-CS* представляет собой дискретное распределение значений $\mathcal{A}$. Так как *Distr2-CS* предназначен для предсказания значения $\mathcal{A}$ для некоторой вершины x, логично потребовать от него быть независимым от значения y_x. Другими словами, вектор для вершины x будет одинаковым, все зависимости от того, известно ли значение атрибута y_x для этой вершины или нет. Для этой цели введём модифицированные значения s и cs, введённые в формулах (2.2) и (2.3), таким образом, чтобы не использовалась информация об y_x.

Определим для вершины v относительную специфичность $\hat{s}(v|x)$, игнорирующую вершину x:

$$\hat{s}(v|x) = KL\big(d(N_v \backslash \{x\})\big|\big|d(Y)\big) \tag{3.3}$$

Аналогичным образом определим относительную специфичность контекста $\hat{cs}(z|x)$ для вершин z и x относительно вершины x:

$$\hat{cs}(z|x) = \sum_{v \in N_x \cap N_z} \hat{s}(v|x) \tag{3.4}$$

Теперь определим *Distr2-CS*. Вектор представляет собой распределение (англ.**Distr**ibution) значений $\mathcal{A}$ для пользователей из **2**-шаговой окрестности N_x^2), построено с применением *CS*-гипотезы. Каждая компонента вектора $Distr2\text{-}CS(x) \in \mathbb{R}^{|A|}$ определяется следующим образом:

$$Distr2\text{-}CS(x)_i = \frac{\sum\limits_{z \in N_x^2 \cap Y} \mathbb{1}_{y_z = a_i} \cdot \hat{cs}(z|x)}{Norm} \tag{3.5}$$

$Norm$ – нормализационная константа, такая что $Distr2\text{-}CS(x)$ удовлетворяет условию $\sum_{i \in 1}^{|A|} Distr2\text{-}CS(x)_i = 1$. Если все компоненты равны в числителе равны 0, определим $Distr2\text{-}CS(x) \equiv \vec{0}$. Стоит отметить, что на практике такой случай не встречался.

Объясним *Distr2-CS* на примере, изображённом на Рисунке 3.1.

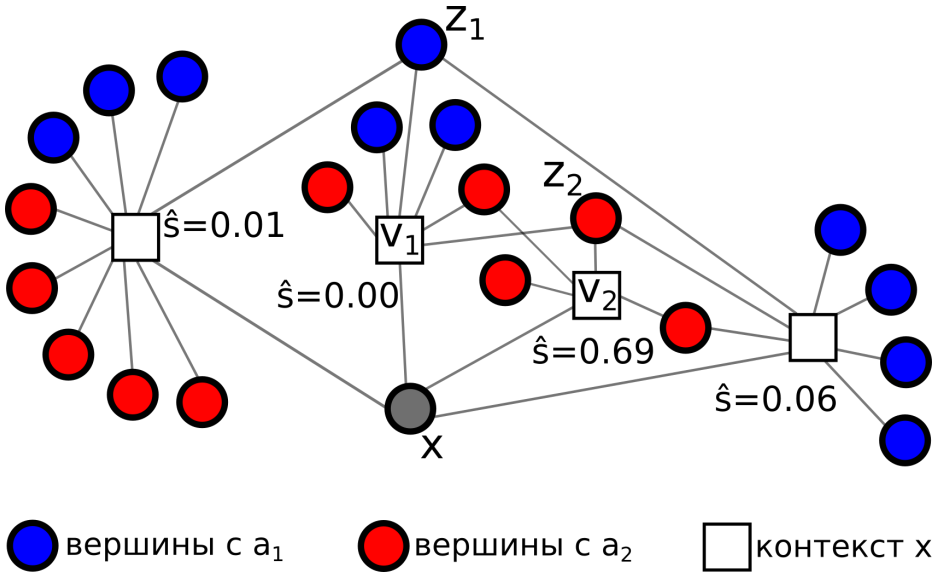

Рисунок 3.1 — Пример для объяснения признаков Distr2-CS, значения относительной специфичности контекста

На рисунке изображено подмножество графа, используемое для вычисления *Distr2-CS* для заданной вершины x. Вершина x обозначена серым цветом. Синие и красные вершины представляют двухшаговую окрестность N_x^2 вершины x, размеченные значениями a_1 и a_2, соответственно. Отдельно из них выделены вершины z_1 и z_2. Белые квадраты представляют вершины из первой

окрестности N_x^1 вершины x, или контекстные вершины. Для каждой контекстной вершины v изображено значение относительной специфичности контекста $\hat{s}(v|x)$. Распределение генеральной совокупности $d(Y) = [0.5, 0.5]^T$. Относительная специфичность $\hat{s}(v_1|x)$ вершины v_1, соединённой с 3 синими и 3 красными вершинами в таком случае равна 0. Относительная специфичность $\hat{s}(v_2|x)$ вершины v_2, соединённой только с красными вершинами достигает максимального значения. Относительная специфичность контекста $\hat{cs}(z_2|x)$ для вершин z_2 и x равна $0.69+0.06+0.0 = 0.75$. Относительная специфичность контекста $\hat{cs}(z_1|x)$ для вершин z_1 и x равна $0.01 + 0.0 + 0.06 = 0.07$. Стоит отметить, обе вершины z_1 и z_2 имеют максимальный в данном примере размер общего с вершиной x контекста, равный 3, хотя их относительная специфичность относительно x существенно отличается. Общая связь вершин x и z_2 с вершиной v_2 имеет большую значимость, чем общая связь вершин x и z_1 с вершиной v_1.

Применив классификатор или регрессор XGBoost [101] для данных признаков получим новый метод, который обозначим *Distr2-CS-XGB*. Суть метод заключается в следующем. Сначала для всех вершин, представляющих пользователей, вычисляются признаки *Distr2-CS*. Для этого используется алгоритм 6. Размеченные пользователи Y вместе со значениями атрибута и векторами *Distr2-CS* используются для обучения модели XGBoost. После чего данная модель используется для предсказания значений атрибута по векторам *Distr2-CS* для остальных пользователей.

3.1.4 Distr2-CS+DW[n]: конкатенация признаков

Описанные выше признаки могут комбинироваться с другими существующими признаками. В данной работе рассматривается конкатенация векторов *Distr2-CS* с векторными представлениями вершин графа, полученными методами DeepWalk (обозначим эти признаки как *DW[n]*) и сингулярным разложением матрицы смежности (обозначим эти признаки *SVD[n]*), где n – размерность соответствующих векторов. Соответствующие методы обозначим как *Distr2-CS+DW[n]-XGB* и *Distr2-CS+SVD[n]-XGB*.

Вход : $G = (V,E)$ – граф, y_x – значения атрибута
Выход: $Distr2\text{-}CS : V \to \mathbb{R}$ – метки вершин

$gen := \vec{0}$; // $gen \in \mathbb{R}^{|A|}$, генеральная совокупность
for $x \in V$ **do**
  $gen[y_x] := gen[y_x] + 1$;
  $d[x] := \vec{0}$; // $d[x] \in \mathbb{R}^{|A|}$
  for $v \in N_x \cap Y$ **do**
    $d[x][y_x] := d[x][y_x] + 1$;
  end
end
$gen :=$ normalize(gen) ;
// вычисление Distr2-CS
for $x \in V$ **do**
  $d2[x] := \vec{0}$; // $d[x] \in \mathbb{R}^{|A|}$, инициализация *Distr2-CS* для x
  for $v \in N_x$ **do**
    $d_{tmp} := d[v]$;
    if $\exists y_x = a_i$ **then**
      $d_{tmp}[y_x] := d_{tmp}[y_x] - 1$; // игнорируем метку вершины x
    end
    $\hat{s} := KL(d_{tmp}||gen)$;
    $d2[x] := d2[x] + d_{tmp} \cdot \hat{s}$;
  end
  $d2[x] :=$ normalize($d2[x]$) ;
end
return $d2$;

Алгоритм 6: Вычисление признаков Distr2-CS для всех вершин графа

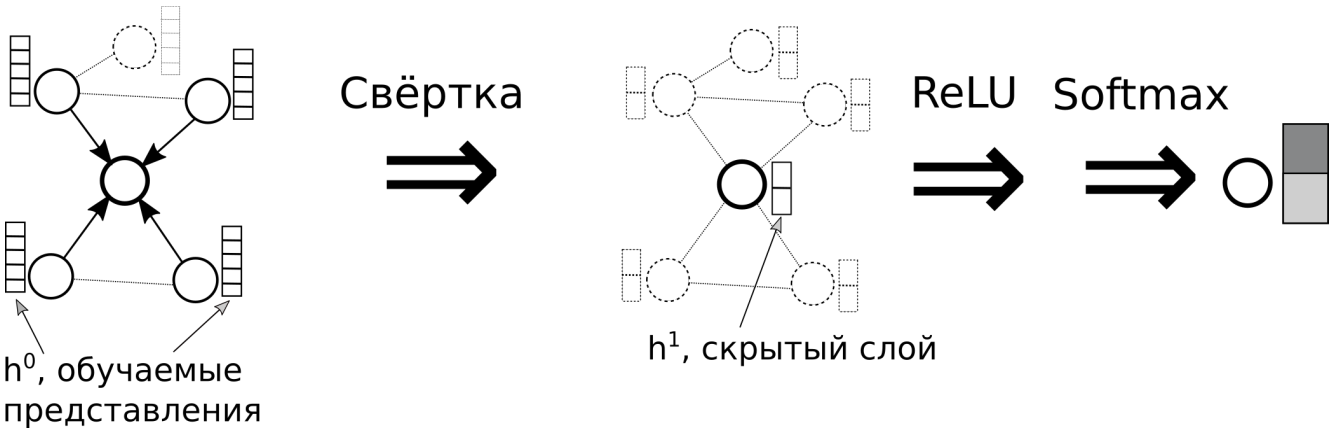

Рисунок 3.2 — Схема графовой нейронной сети GConv-CS[n]

3.1.5 GConv-CS: регуляризация свёрточной графовой нейронной сетей

Метод *GConv-CS[n]* представляет собой свёрточную графовую нейронную сеть со следующей структурой. Для каждой вершины рассматриваются обучаемые в рамках модели векторные представления $h_z^0 \in \mathbb{R}^n$, где n – размер векторов, гиперпараметр модели. Затем для каждой вершины применяется свёртка с обучаемыми параметрами из векторных представлений соседних вершин в вектора размера $|A|$. Свёртка представляет собой следующее преобразование:

$$h_x^1 = b + \sum_{z \in N_x} \frac{1}{\sqrt{|N_x| \cdot |N_z|}} W \cdot h_z^0 \tag{3.6}$$

Здесь $W \in \mathbb{R}^{n \times |A|}$ и $b \in \mathbb{R}^{|A|}$ – обучаемые параметры свёртки, $h_x^1 \in \mathbb{R}^{|A|}$ – выходной вектор свёртки до применения активации. К h_x^1 применяется активация ReLU (линейный выпрямитель), затем Softmax.

Функция потерь основана на логарифмической функции правдоподобия: $L = -\log\big(p_x[y_x]\big)$, где p_x – выход softmax, $p_x[y_x]$ – компонент, соответствующий правильному ответу. Дополнительно к функции потерь добавляется регуляризация $MSE(l2(x), s(x))$. Цель регуляризации – добиться, чтобы $l2$ норма векторных представлений соответствовала специфичности контекста вершин. Данный метод применятся только для задач классификации, т.е. для предсказания значений атрибутов пол и род деятельности в рассматриваемых наборах данных.

3.2 Оценка вычислительной сложности методов

Лемма 1 Вычислительная сложность методов *LP-CS*, *LP-Gen*, алгоритма вычисления признаков *Distr2-CS* для всех вершин графа составляет $O\Big(\ |V|\ +\ |E|\ \Big)$.

Доказательство. Методы LP-CS, LP-CS-Gen и алгоритм вычисления признаков Distr2-CS для всех вершин похожи по своей структуре. Во всех трёх случаях необходимо вычислить распределение значений атрибута для всех вершин, для чего необходимо $O\big(|V|\big)$ шагов алгоритма. Далее выполняются два прохода по всем соседям всех вершин. В каждом проходе каждое ребро (x,z) графа задействовано два раза: при обращении к соседям вершин x и z. Таким образом, общая вычислительная сложность методов LP-CS, LP-CS-Gen составляет $O\Big(|V| + |E|\Big)$. Аналогичная вычислительная сложность для вычисления признаков Distr2-CS. □

Теорема 2 Пусть n – размер векторных представлений вершин. Тогда вычислительная сложность методов составляет:

- *LP-CS* и *LP-Gen* – $O\Big(|V| + |E|\Big)$;
- *Distr2-CS-XGB* – $O\Big(|V|\log|V| + |E|\Big)$;
- *Distr2-CS+DW[n]-XGB* – $O\Big(n|V|\log|V| + |E|\Big)$;
- *GConv-CS[n]* – $O\Big(n|V| + n|E|)\Big)$.

Доказательство.

Вычислительная сложность *LP-CS* и *LP-Gen* доказана в лемме 1.

Вычислительная сложность XGBoost [101] в случае плотных признаковых представлений объектов константного размера $|A|$ составляет $O\big(|V|\log|V|\big)$. Для метода *Distr2-CS-XGB* необходимо вычислить признаки *Distr2-CS* для всех вершин, представляющих пользователей, затем применить XGBoost. Применяя лемму 1 получаем, что общая вычислительная сложность метода *Distr2-CS-XGB* составляет $O\Big(|V|\log|V| + |E|\Big)$.

Для *Distr2-CS+DW[n]-XGB* необходимо вычислить признаки Distr2-CS и DeelWalk для всех вершин. Из леммы 1 сложность вычисления Distr2-CS составляет $O\big(|V| + |E|\big)$. Вычислительная сложность DeepWalk [57; 81] составляет $O\big(n|V|\big)$. К объединённым признакам применяется алгоритм XGBoost. Вычислительная сложность XGBoost [101] в случае плотных признаковых

представлений объектов размера $|A| + n$, где $|A|$ – константа, $n \to \infty$, составляет $O\big(n|V|\log|V|\big)$. Таким образом, общая вычислительная сложность метода *Distr2-CS+DW[n]-XGB* составляет $O\Big(n|V|\log|V| + |E|\Big)$.

При обучении метода *GConv-CS* необходимо применить свёртку для графа. Это занимает $O\big(n|E|\big)$ шагов, так как функция свёртки применяется к векторам размера n для всех соседей каждой вершины, для чего необходимы проходы по всем рёбрам. Вычисление потерь в методе GConv, т.е. без учёта регуляризации, занимает $O\big(|V|\big)$ шагов. При использовании регуляризации необходимо вычислить значения специфичности контекста для всех вершин графа. При этом вычисляется генеральная совокупность за $O\big(|V|\big)$ шагов и просматриваются все соседи всех вершин за $O\big(|E|\big)$ шагов. Кроме того, для регуляриации необходимо вычислять нормы обучаемых векторов, что занимает $O\big(n|V|\big)$ шагов. Всего для вычисления потерь необходимо $O\big(n|V| + |E|\big)$ шагов. Необходимо обновлять обучаемые параметры свёртки, для чего необходимо $O\big(n\big)$ шагов, а также обучаемые входные векторы для всех вершин, для чего необходимо $O\big(n|V|\big)$. Таким образом, общая вычислительная сложность метода *GConv-CS[n]* составляет $O\Big(n|V| + n|E|\Big)$. □

3.3 Обсуждение

Обсудим некоторые особенности предлагаемых признаков *Distr2-CS* по сравнению со статическими представлениями вершин графа, особенности и ограничения разработанных методов.

Статические векторные представления вершин графа, такие как DeepWalk и LINE обычно вычисляются в режиме «оффлайн», что означает, что сразу вычисляются все векторные представления всех вершин, для чего необходим весь граф целиком. Признаки *Distr2-CS* могут быть вычислены в «онлайн» режиме для одной заданной вершины, причём для этого не нужен весь граф, достаточно лишь второй окрестности заданной вершины. Эта особенность может быть полезна на практике, так реальные социальные графы содержат сотни миллионов узлов, не всегда в наличии имеется достаточное количество ресурсов для хранения в основной памяти и обработки графов такого размера.

Однако статические представления имеют преимущество по сравнению в Distr2-CS. При решении нескольких задач предсказания различных атрибутов, векторы Distr2-CS необходимо пересчитывать для каждой задачи. Статические представления не зависят от значений атрибутов, поэтому достаточно их вычислить один раз для социального графа.

Стоит также отметить, что демографические характеристики являются свойством пользователей. Социальный граф может содержать в том числе и вершины других типов, например, публичные страницы организаций, событий и т.д. Все разработанные методы могут быть применены к двудольным графам, например пользователи-лайки, содержащий информацию об отметках «мне нравится», которые пользователи ставят для различных объектов. Методы, использующие только значения атрибутов первую окрестности, не применимы к таким графам.

Методы LP-CS, LP-CS-Gen, методы на основе Distr2-CS, и метод GConv-CS используют социальный граф, собранный для множества целевых вершин согласно алгоритму 3, то есть используются связи для каждой их целевых вершин. Информация о наличии рёбер между вершинами на первой окрестности от целевых вершин никак не используется этими методами (в отличие, например, от методов, основанных на статических векторных представлениях).

Метод GConv-CS[n], основанные на графовых нейронных сетях, применим только для задач классификации и не предназначены для решения задач регрессии. LP-CS-Gen и методы, основанные на признаках Distr2-CS также в теории не предназначены для задач регрессии. Однако рассматриваемые в диссертационной работе наборы данных и атрибуты, предсказание значений которых сводится к регрессии (доход и возраст) обладают особенностью. Множество различных значений, которые принимают эти атрибуты, существенно меньше, чем количество примеров, то есть размеченных пользователей (см. таблицу 3). Это связано с тем, что возраст в наборе данных представлен с точностью до года, а для определения референсных значений дохода использовались средние значения для соответствующего рода деятельности. Таким образом, в случае LP-CS-Gen метка выбирается среди относительно небольшого (по сравнению с размером набора данных) количества различных значений, Distr2-CS получается не таким разреженным, как если бы все значения атрибуты были различными. При попытке применить к таким данным разработанные методы, было выявлено, что на LP-CS-Gen и GConv-CS не показывают при-

емлемого качества, а методы, основанные на признаках Distr2-CS, показывают качество не хуже, а в некоторых случаях и выше, чем другие методы, применимые к задачам регрессии. Таким образом, экспериментально показано, что методы, основанные на использовании признаков Distr2-CS, применимы с случае $|A| \ll |V|$.

3.4 Экспериментальное сравнение методов

В разделе описываются результаты экспериментального сравнения разработанных методов предсказания значений демографических атрибутов на основе специфичности контекста и аналогичных базовых методов, не использующих данное свойство. Вначале опишем эти базовые методы.

LP[1] и *LP[2]* – синхронный алгоритм распространения меток с одной и двумя итерациями, соответственно. Метка представляет собой значение атрибута, на каждом шаге алгоритма новое значение метки для вершины вычисляется как самое частое значение среди меток соседей (для атрибутов пол и род деятельности), или среднее значение метки (для атрибутов возраст и доход).

Distr2-XGB модификация авторского метода *Dist2-CS-XGB*, игнорирующая значения специфичности контекста. Вместо специфичности общего контекста используется размер общего контекста:

$$Distr2(x)_i = \frac{\sum\limits_{z \in N_x^2 \cap Y} \mathbb{1}_{y_z = a_i} |N_x \cap N_z|}{Norm} \tag{3.7}$$

DW[n]-XGB заключается в построении векторных представлений, полученных методом DeepWalk, и применении к данным представлениям классификатора или регрессора XGBoost.

GConv[n] аналогичен методу *GConv-CS[n]*. Отличие в том, что в *GConv[n]* не применятся регуляризация, норма $l2$ не связывается со специфичностью контекста вершины. Данный метод также применяется только для задач классификации.

Опишем процесс экспериментального сравнения методов. Всего имеется 9 различных комбинаций (набор данных, атрибут). 5 из них – задачи классификации (род деятельности, пол), оставшиеся 4 – задачи регрессии (возраст,

доход). Для каждой из 9 комбинаций выполнялась следующая последовательность действий. Множество размеченных пользователей Y случайным образом разбивалось на обучающую (80%) и проверочную (20%) части. Эксперименты были проведены также и для других пропорций разбиения на тренировочные и тестовые данные, их результаты с описываются в приложении Б. Каждый из методов был применен к обучающей выборке для построения модели предсказания, затем полученная модель была использована для оценки качества на проверочной части. Для задач классификации использовались значения F1-меры с микро- и макроусреднением. Для задач регрессии использовались метрики R2 и среднеквадратичная ошибка (MAE). Данный процесс повторялся по 30 раз для каждой комбинации (набор данных, атрибут). После чего были подсчитаны средние значения метрик и доверительные интервалы, полученные с использованием t-распределения Стьюдента [103], с уровнем доверия $p = 0.95$. Для набора данных pokec в качестве размера векторных представлений в методах DW[n]+XGB, Distr2-CS+DW[n]-XBG, GConv[n], GConv-CS[n] использовалось значение $n = 128$, используемое ранее в работе [82]. Для остальных наборов данных использовалось значение $n = 32$, используемое ранее в работе [85].

Сокращённые обозначения методов, используемые в графиках и таблицах:

– D2 – Distr2-XGB;
– D2-CS – Distr2-CS-XGB;
– DW[n] – DW[n]-XGB;
– DW[n]+D2-CS – Distr2-CS+DW[n]-XGB.

На рисунках 3.5 - 3.11 представлены результаты экспериментального сравнения методов. Из рисунков видно, что методы, основанные на специфичности контекста *LP-CS*, *Distr2-CS-XGB*, *DW[n]+Distr2-CS-XGB*, *GConv-CS[n]* показывают более высокое качество по сравнению с методами *LP[2]*, *Distr2-XGB*, *DW[n]-XGB*, *GConv[n]* соответственно. *LP[1]* показывает сравнимые с лучшими методами результаты предсказания рода деятельности и возраста в наборе данных vk1. Это свидетельствует о выраженной гомофилии в этом наборе данных для этих атрибутов. Это связано как с природой этих атрибутов, так и со способом сбора социального графа. Набор данных представляет собой сообщество пользователей, объединённых вокруг МГУ им. М.В. Ломоносова: студентов и выпускников. Многие пользователи знают друг друга, причём знакомство связано как правило с тем, что люди учатся в одной группе или на одном факультете,

работают в одной сфере и т.д. Сбор значений рода деятельности производился при помощи аннотаторов, которые отмечали свой род деятельности и род деятельности своих друзей, у которых они его знали.

В таблице 13 отображены агрегированные результаты. Рассматриваются только предложенные методы: LP-CS, LP-CS-Gen, D2-CS, DW[n]+D2-CS, GConv-CS[n]. Опишем способ вычисления агрегированной метрики для каждой пары (набор данных, атрибут). Рассмотрим одну вычисленную метрику, найдём метод с наибольшим средним значением метрики. Добавим методы, доверительные интервалы которых пересекаются с доверительным интервалом этого метода. Разделим 1 балл равномерно между полученными методами. Для второй метрики разделим 1 балл между методами аналогичным способом. таким образом, для каждой пары (набор данных, атрибут) оценка в 2 балла делится между методами. Затем отдельно для задач классификации и задач регрессии вычислим суммарный балл для каждого метода. Из агрегированных результатов в таблице 13 можно сделать вывод, что для задач классификации наибольшее качество показывает метод GConv-CS[n], а для задач регрессии – метод Distr2-CS+DW[n]-XGB.

Время работы методов представлено в таблице 14. Для метода Distr2-CS+DW[n]-XGB отдельно представлено время работы метода при заданных векторных представлениях DeepWalk (без вычисления представлений) и время вычисления векторны представлений. Время работы методов замерялось на виртуальной машине с процессором Intel(R) Core(TM) i7-9700F с частотой 3.0ГГц. Под виртуальную машину было выделено 8 ядер и 24 гигабайта основной памяти. Для вычисления векторных представлений DeepWalk для графа pokec использовалась виртуальна машина с 64 ГБ основной памяти, 8 ядрами ЦП, созданная в облаке ИСП РАН, так как 24 ГБ памяти не хватило для вычисления. Стоит отметить, что самым долгим этапом работы является вычисление векторных представлений вершин графа DeepWalk. Однако эти представления достаточно вычислить один раз для каждого графа, остальные же методы необходимо запускать с нуля при различных атрибутах и при различных разбиениях на тренировочную и тестовую части, в рамках описанного экспериментального сравнения.

Итак, результаты экспериментального сравнения методов показали, что специфичность контекста является важным признаком, позволяющим повы-

Таблица 13 — Агрегированная метрика сравнения качества разработанных методов

набор данных	атрибут	LP-CS	LP-CS-Gen	D2-CS	DW[n]+D2-CS	GConv-CS[n]
twitter	род деятельности					2.0
vk1	род деятельности		0.25	0.25	0.75	0.75
vk1	пол					2.0
vk2	пол				1.0	1.0
pokec	пол				2.0	
twitter	доход	0.33		0.83	0.83	
vk1	возраст					
vk2	возраст			1.0	1.0	
pokec	возраст			0.5	1.5	
Классификация		0	0.25	0.25	3.75	**5.75**
Регрессия		0.33		2.33	**3.33**	

сить качество предсказания значений стационарных демографических атрибутов пользователей социальных сетей.

3.5 Рекомендации к использованию разработанных методов

На основе теоретической оценки вычислительной сложности и экспериментального сравнения разработанных методов по качеству и скорости представлены рекомендации к использованию разработанных методов.

Рассмотрим задачу классификации. Если необходимо быстрое решение с приемлемым качеством, то рекомендуется использовать метод LP-Gen-CS. Если необходимо максимально качественное решение, то рекомендуется использовать GConv-CS[n]. Значение n необходимо подбирать эмпирически, при бóльшем количестве вершин в графе рекомендуется использовать бóльшее значение n. Если необходим компромисс между скоростью и качеством работы, предлагается вос-

Таблица 14 — Время работы методов, чч:мм:сс

граф	атрибут	Вычисление DW[n]	DW[n]+D2-CS (без DW)	D2-CS	LP-CS	LP-CS-Gen	GConv-CS[n]
twitter	род деят.	00:05:48	00:00:17	00:00:14	00:00:06	00:00:06	00:00:56
	доход		00:00:41	00:00:39	00:00:08		
vk1	род деят.	00:00:30	00:00:02	00:00:02	00:00:01	00:00:01	00:00:06
	пол		00:00:01	00:00:01	00:00:01	00:00:01	00:00:04
	возраст		00:00:01	00:00:01	00:00:01		
vk2	пол	00:10:23	00:00:08	00:00:07	00:00:09	00:00:09	00:01:02
	возраст		00:00:30	00:00:29	00:00:11		
pokec	пол	05:40:00	00:12:14	00:03:54	00:04:18	00:04:06	01:10:57
	возраст		00:33:40	00:25:30	00:04:31		

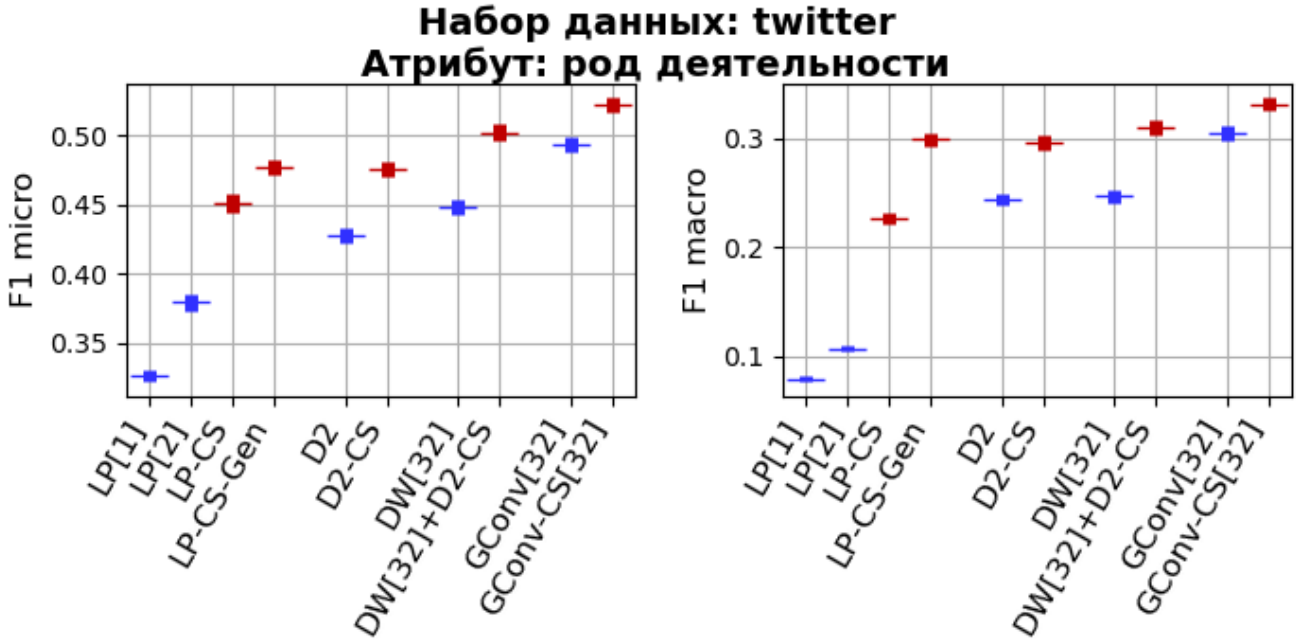

Рисунок 3.3 — Результаты экспериментального сравнения методов для набора данных twitter; атрибут: род деятельности

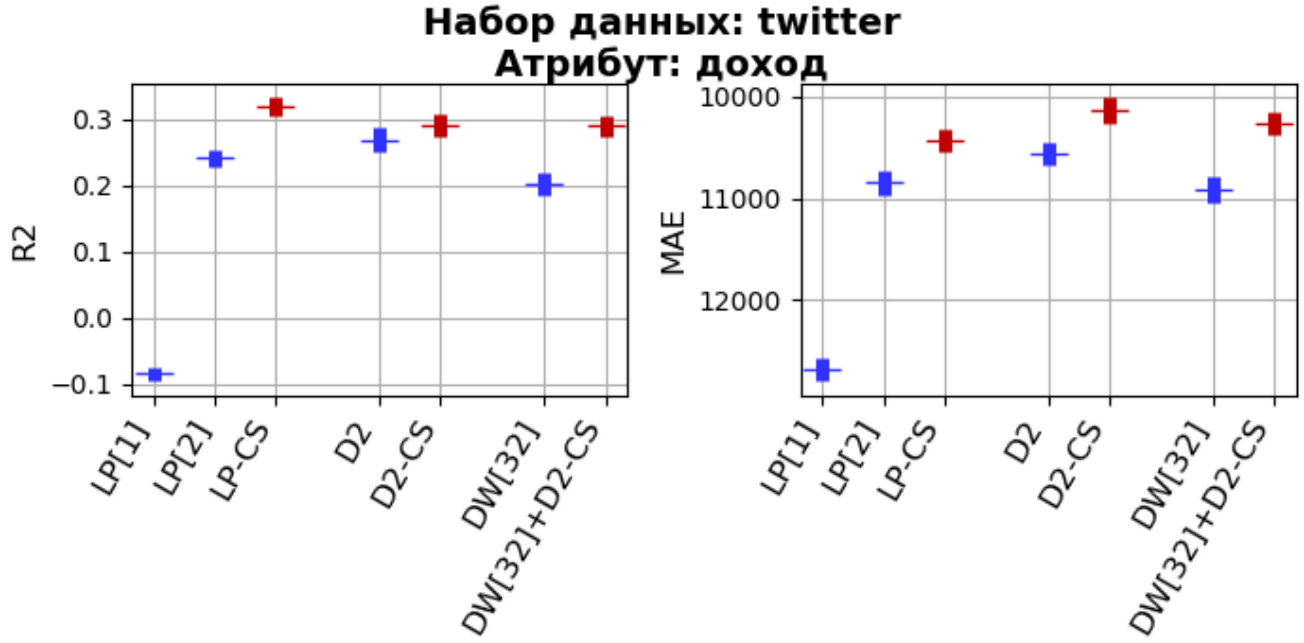

Рисунок 3.4 — Результаты экспериментального сравнения методов для набора данных twitter; атрибут: доход

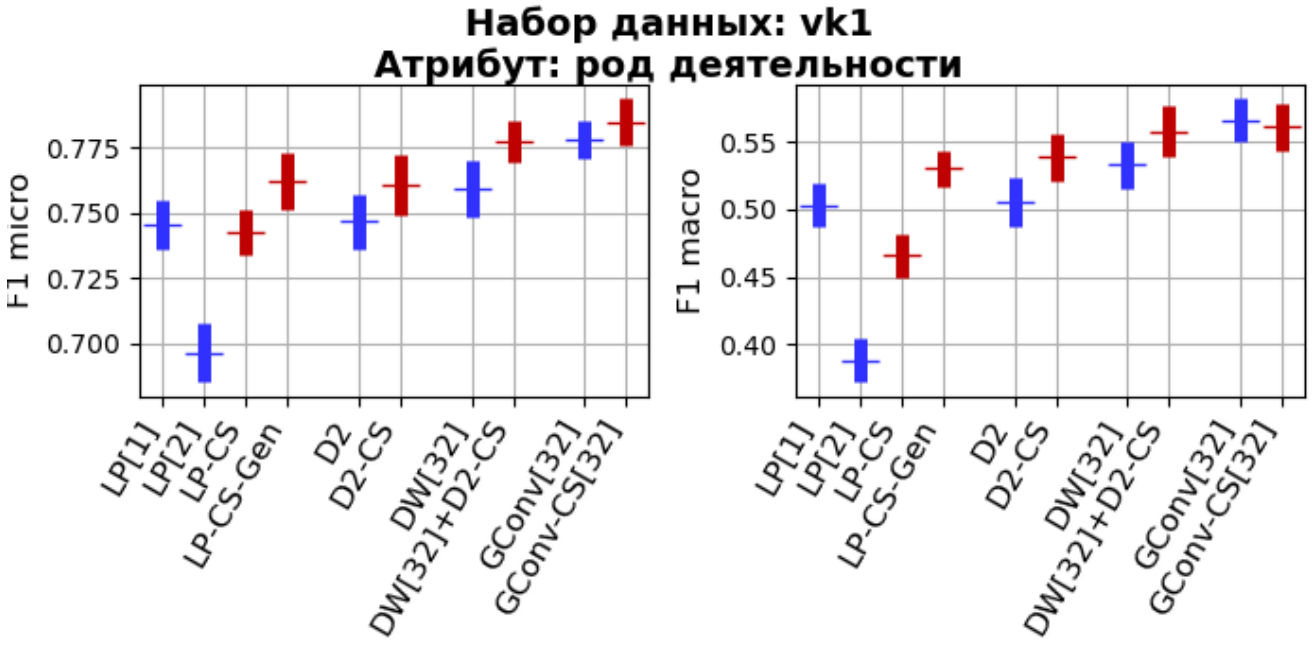

Рисунок 3.5 — Результаты экспериментального сравнения методов для набора данных vk1; атрибут: род деятельности

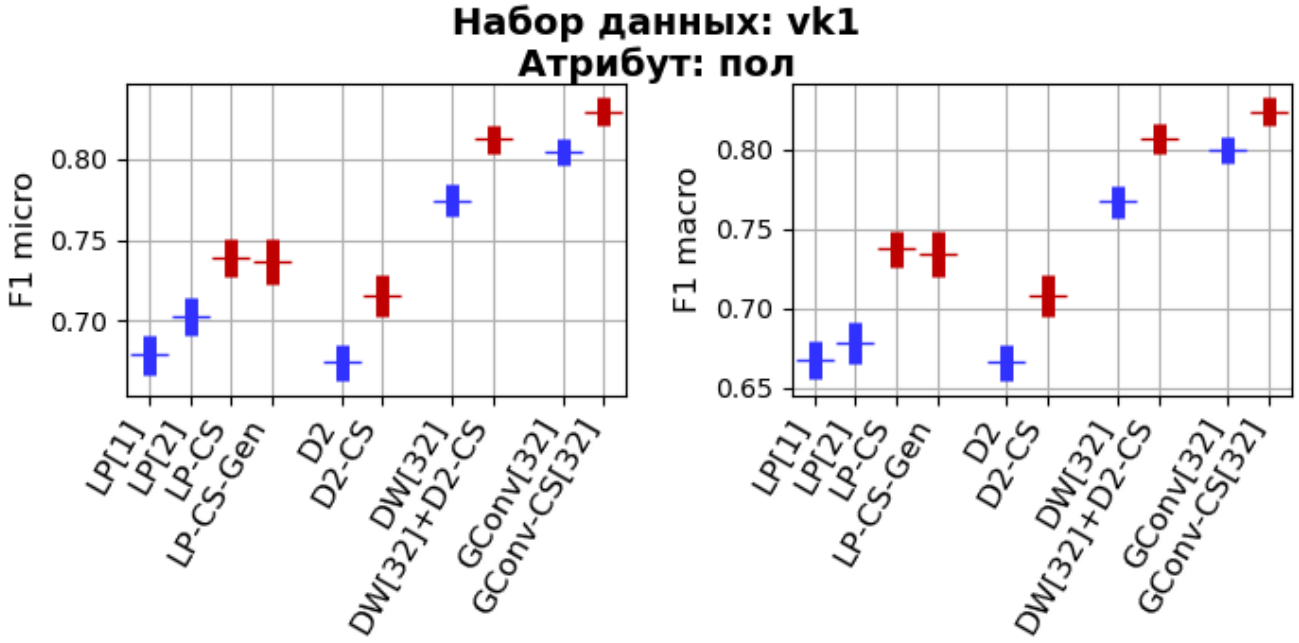

Рисунок 3.6 — Результаты экспериментального сравнения методов для набора данных vk1; атрибут: пол

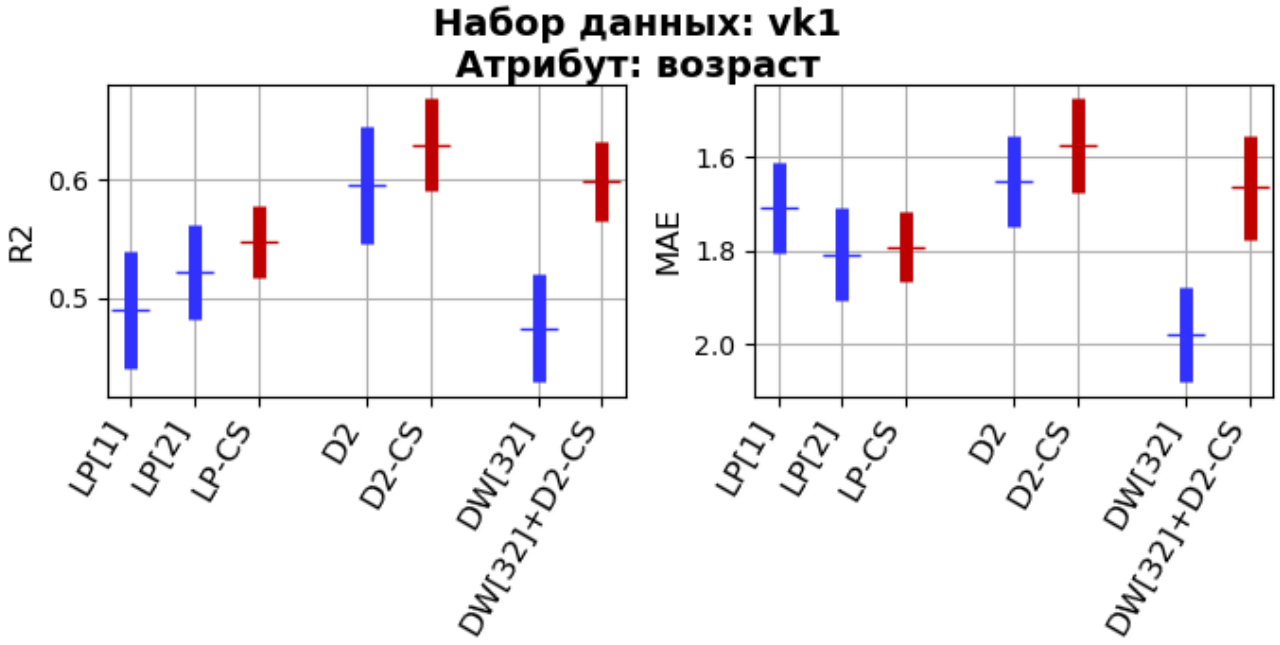

Рисунок 3.7 — Результаты экспериментального сравнения методов для набора данных vk1; атрибут: возраст

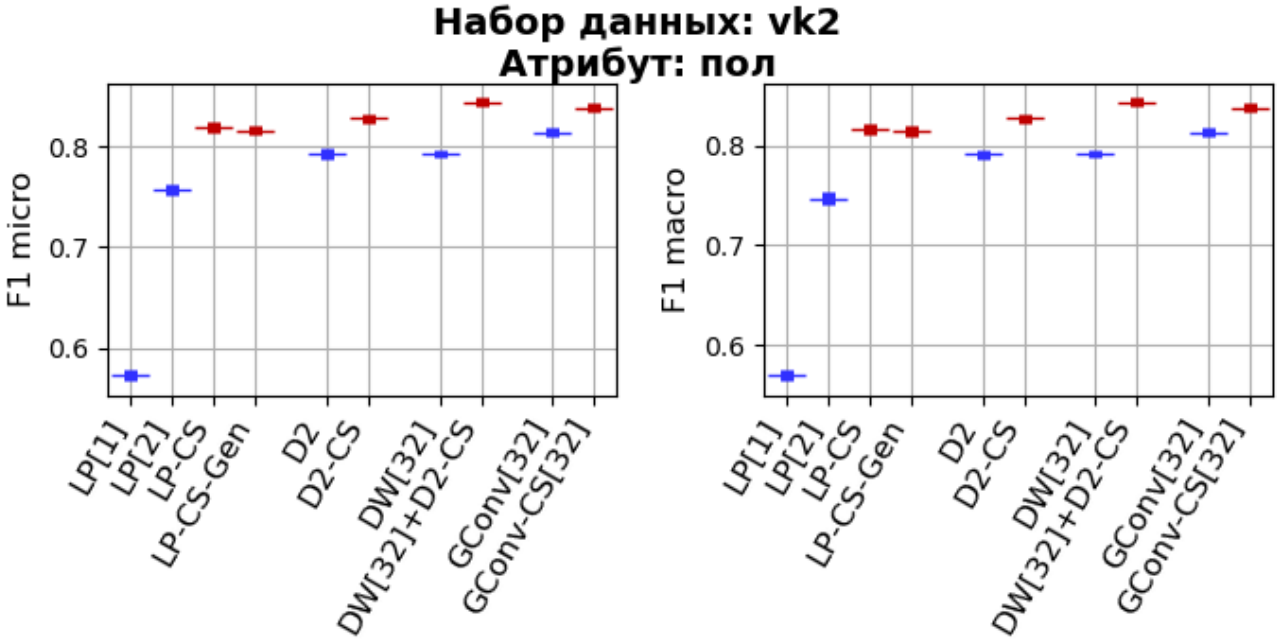

Рисунок 3.8 — Результаты экспериментального сравнения методов для набора данных vk2; атрибут: пол

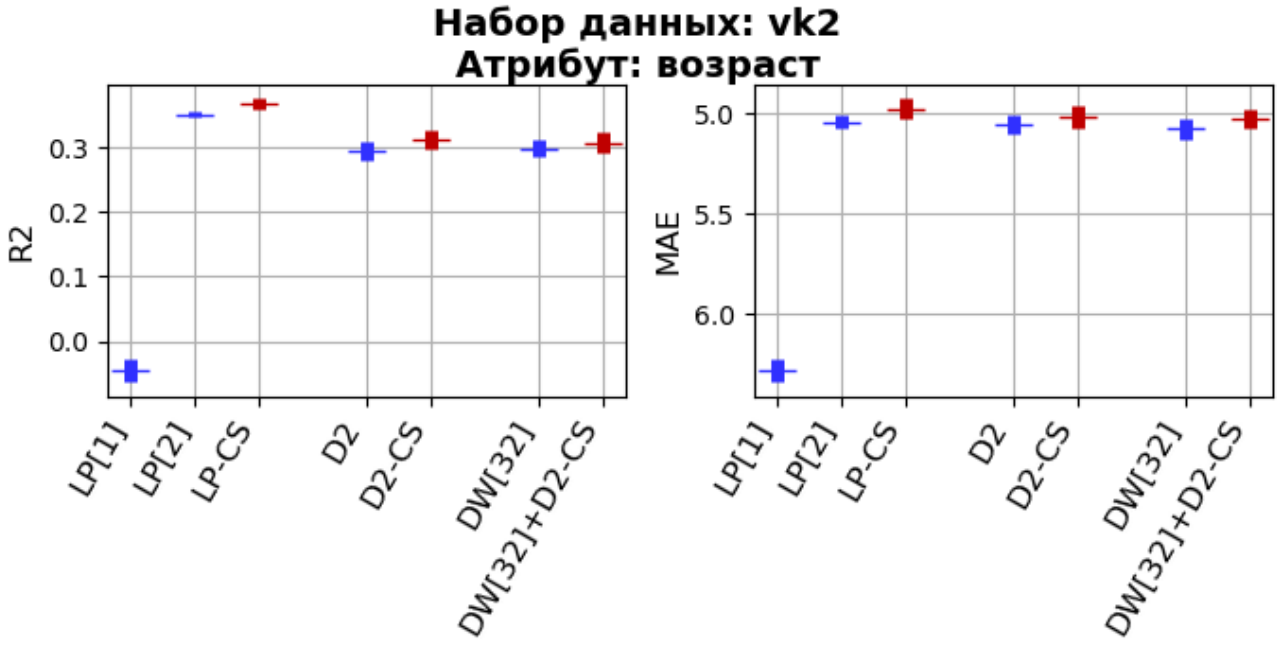

Рисунок 3.9 — Результаты экспериментального сравнения методов для набора данных vk2; атрибут: возраст

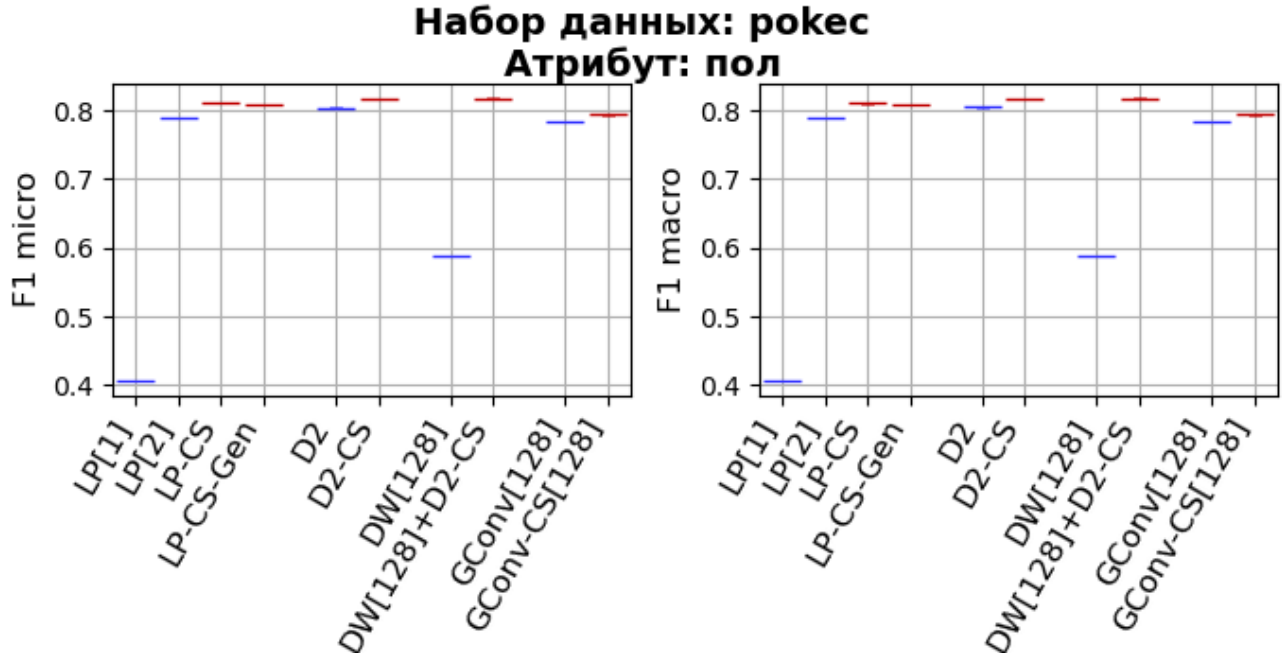

Рисунок 3.10 — Результаты экспериментального сравнения методов для набора данных pokec; атрибут: пол

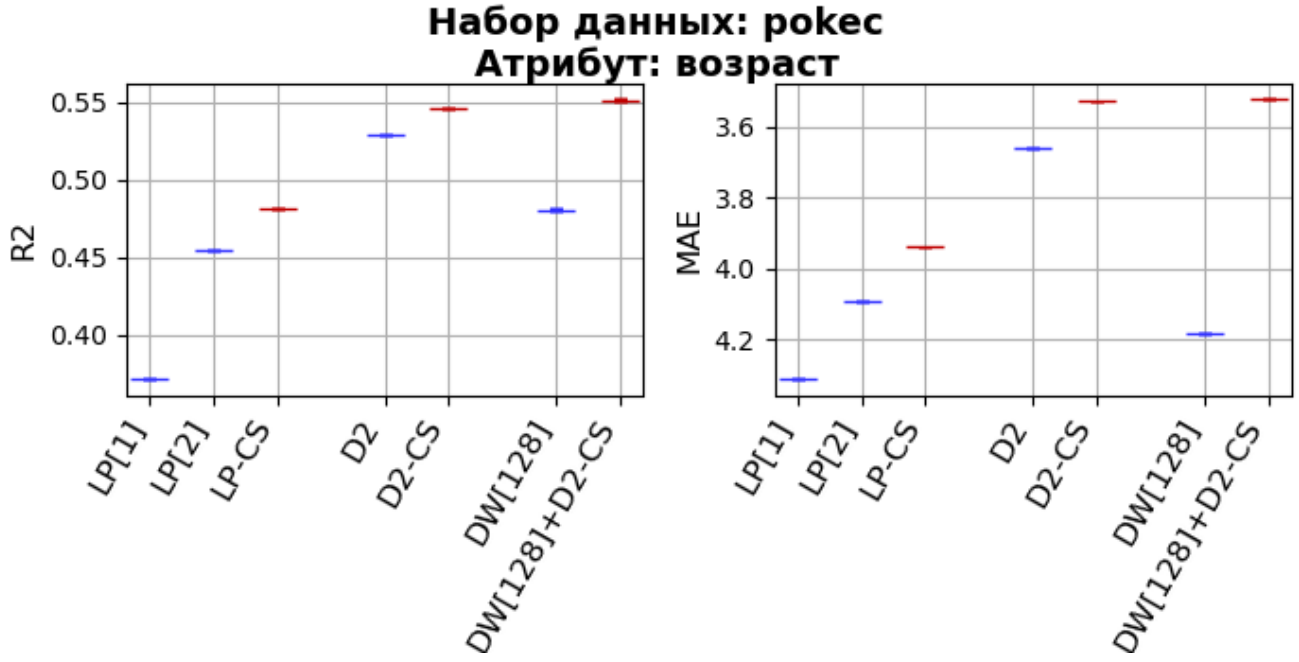

Рисунок 3.11 — Результаты экспериментального сравнения методов для набора данных pokec; атрибут: возраст

Таблица 15 — Свойства разработанных методов предсказания значений атрибутов

Метод	Скорость	Применимость		Качество	
		Класс.	Регр.	Класс.	Регр.
LP-CS	*****	да	да	**	**
LP-CS-Gen	*****	да	нет	***	
Distr2-CS-XGB	****	да	$\|A\| \ll \|V\|$	***	***
Distr2-CS+DW[n]-XGB	*** / *	да	$\|A\| \ll \|V\|$	****	****
GConv-CS	**	да	нет	*****	

пользоваться методам Distr2-CS-XGB. Метод DW[n]Distr2-CS-XGB показывает более высокое качество, чем Distr2-CS-XGB, однако требует вычисления векторных представлений DeepWalk, которое занимает существенную часть времени. Однако если эти признаки вычислены заранее для графа, нет необходимости вычислять их заново, так как они не зависят от атрибута и его значений для вершин графа. Поэтому в таблице 15 в графе скорость обозначено «*** / *», что означает среднюю скорость работы, если статические векторные представления DeepWalk уже вычислены заранее, и самую медленную скорость, если готовых векторных представлений нет, и их нужно вычислять.

Не все из представленных методов применимы для задач регрессии. Рекомендуется использовать метод LP-CS, который показывает как высокую скорость, так и относительно высокое качество. В некоторых случаях более высокое качество достигается с использованием методов Distr2-CS или Distr2-CS+DW. Однако в случае регрессии эти методы применимы только когда $|A| \ll |V|$. Вычислительная сложность Distr2-CS+DW выше, чем Distr2-CS, но в некоторых случаях Distr2-CS+DW показывал более высокое качество.

3.6 Выводы

В рамках подхода на основе специфичности контекста были предложены методы *LP-CS* и *LP-CS-Gen* для предсказания значений атрибутов пользователей, являющийся вариацией алгоритма распространения меток. На основе специфичности контекста предложены признаки для представления вершины

графа *Distr2-CS*. Предложены методы *Distr2-CS-XGB* и *Distr2-CS+DW[n]-XGB* для предсказания значений атрибутов пользователей, основанные на признаках *Distr2-CS* и использующие классификатор XGBoost. Предложен метод *GConv-CS[n]* для предсказания значений атрибутов пользователей, основанный на свёрточных графовых нейронных сетях. Сформулирована и доказана теорема о вычислительной сложности предложенных методов.

Проведено экспериментальное сравнение разработанных методов на основе специфичности контекста с аналогичными методами, не использующие специфичность контекста. Результаты экспериментального сравнения показали, что специфичность контекста является важным признаком, позволяющим повысить качество предсказания значений стационарных демографических атрибутов пользователей социальных сетей.

Описание методов и результаты экспериментального сравнения опубликованы в работе [1].

Глава 4. Программная система для предсказания значений демографических атрибутов пользователей социальных сетей

Программная система для предсказания значений демографических атрибутов пользователей социальных сетей по социальному графу и представляет собой фреймворк. В нём реализованы методы предсказания значений атрибутов пользователей по социальному графу. Фреймворк позволяет сравнить качество различных методов предсказания значений демографических атрибутов. Имеется возможность добавить новые методы, признаковые описания вершин графа, настраивать методы оценки качества. Кроме того, система позволяет оформить результаты экспериментального сравнения методов в виде графиков с настраиваемыми цветами, подписями, типами линий и точек. Реализована возможность проанализировать набор данных с целью оценки свойств гомофилии (H), зависимостей между размером общим контекстом и значениями атрибутов (C), зависимости между специфичностью общего контекста и значениями атрибутов (CS). Результаты анализа оформляются в виде графиков.

Исходный код программной системы составляет около 4100 строк на языке Python 3 и 520 строк на языке C++. На рисунке 4.1 изображена диаграмма классов программной системы.

4.1 Реализация методов предсказания значений демографических атрибутов пользователей

Во фреймфорке введены две абстракции, позволяющие реализовать решение задачи предсказания значений демографических атрибутов пользователей: генераторы признаков и методы.

Генератор признаков (Generator) представляет собой абстракцию с двумя операциями. Основная операция, *gen(ids)*, позволяет сгенерировать признаковые векторы для заданного набора вершин. На вход подаётся список идентификаторов вершин, на выходе – список признаковых представлений для заданных вершин. Вторая вспомогательная операция, *set_train(ids)* используется, чтобы сообщить генератору, какие из идентификаторов относятся к

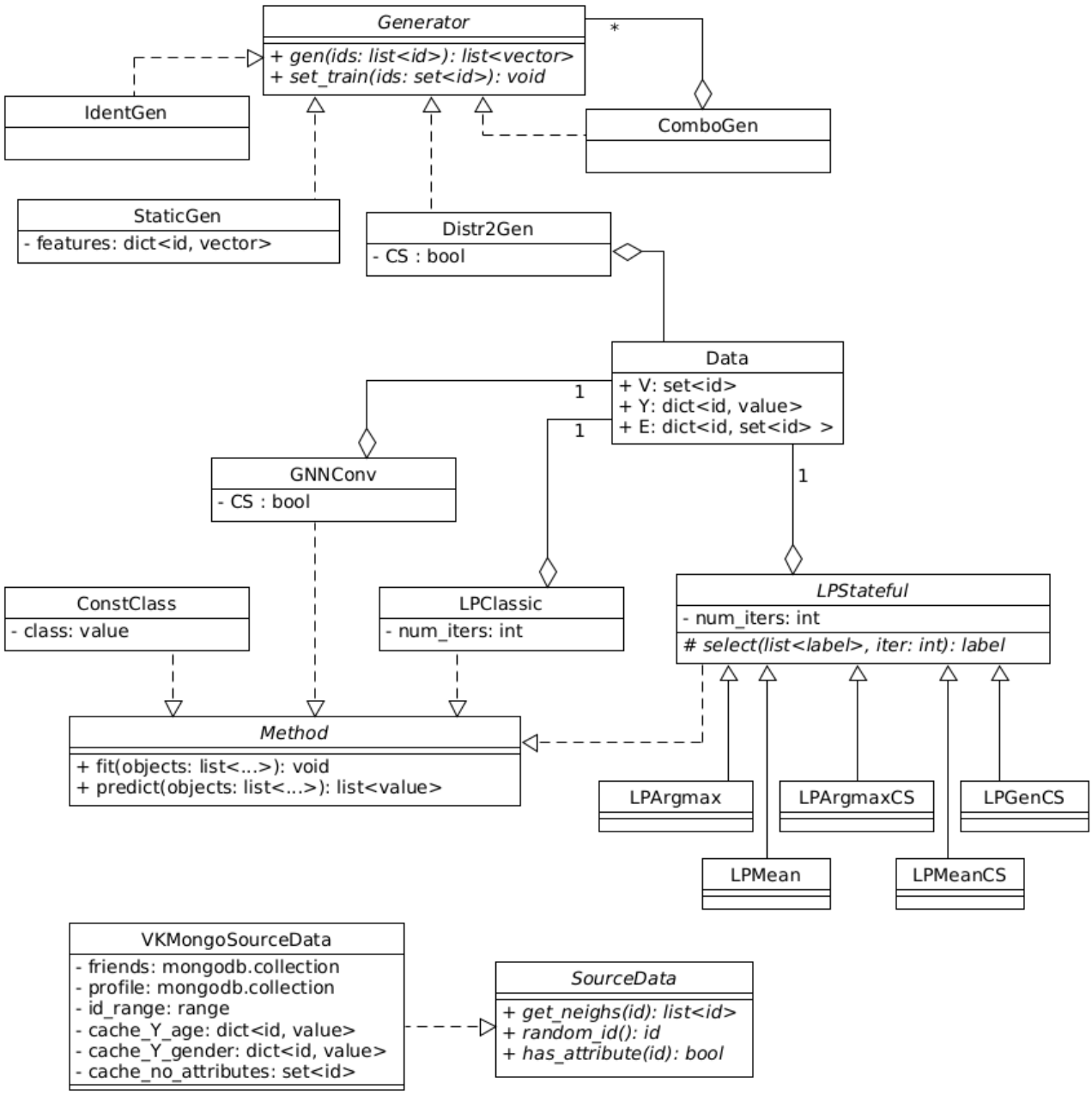

Рисунок 4.1 — Диаграмма классов программной системы

обучающей выборке, и значения атрибутов которых могут быть использованы для генерации признаков. Эта операция введена с целью оптимизации тестирования на одном наборе данных, но с разными множествами вершин с известным значением атрибута. В рамках фреймворка реализованные следующие генераторы признаков:

- InentGen – генератор, который возвращает то, что ему подали на вход, то есть список идентификаторов. Используется в случаях, когда решение не используют признаковые представления вершин.
- StaticGen – генератор, хранящий заранее считанные статические признаки и возвращающий признаковые представления для заданных вершин.
- Distr2Gen – генератор, реализующий вычисление признаков Distr2 и Distr2-CS. Использование или игнорирование специфичности контекста для вычисления признаков является параметром, который задаётся при создании объекта генератора.
- ComboGen – генератор, конкатенирующий признаковые представления вершин графа, полученные с помощью других генераторов. Используемые генераторы задаются при создании объекта.

Методы удовлетворяют интерфейсу методов классификации sklearn [47], включающему операции *fit(X,y)* для обучения и *predict(X)* для предсказания. В рамках фреймворка реализованные следующие методы:

- ConstClass – простой базовый метод, который для всех входящих примеров возвращает одинаковое значение, которое задаётся при создании объетоа.
- LPClassic – асинхронная версия алгоритма распространения меток. Количество итераций задаётся при создании объекта.
- LPStateful – синхронная версия алгоритма распространения меток, абстрактный класс. Реализованы следующие методы, использующие различные способы выбора меток по меткам соседей:
  - LPArgmax – самая частая метка (метод LP[n] для задач классификации).
  - LPMean – среднее значение (метод LP[n] для задач регрессии).
  - LPArgmaxCS – самая частая метка на первом шаге, наиболее специфичная метка на втором шаге (метод LP-CS для задач классификации).

  - LPGenCS – самая выделяющаяся относительного генеральной совокупности метка на первом шаге, наиболее специфичная метка на втором шаге (метод LP-CS-Gen для задач классификации).
  - LPMeanCS – среднее значение на первом шаге, взвешенное на основе специфичности среднее значение на втором шаге (метод LP-CS для задач регрессии).
- GNNConv – метод, реализующий свёрточную графовую нейронную сеть. Использование дополнительная регуляризация на основе специфичности контекста вершин является параметром, задающимся при создании объекта.

4.2 Реализация способов сравнения качества методов

Реализовано два способа сравнения качества методов. Первый способ основан на применении метода скользящего контроля (кросс-валидация). Второй способ заключается в многократном разбиении выборки на обучающую и контрольную в заданной пропорции. По умолчанию для задач классификации измеряются значения F1 меры с микро- и макроусреднением, для задач регрессии измеряются среднеквадратичная ошибка (MAE) и коэффициент детерминации (R2). Фреймворк позволяет использовать другие метрики качества. При оценке качества вычисляются как средние значения, так и доверительные интервалы, полученные с использованием t-распределения Стьюдента [103], с настраиваемым уровнем доверия.

4.3 Реализация визуального оформления результатов

Во фреймворке реализовано построение двух типов графиков, представляющих результаты экспериментального сравнения методов предсказания значений демографический атрибутов. На графике первого типа по оси абсцисс располагаются методы, по оси ординат – качество методов. Средние значения

качества обозначены горизонтальными отрезками, доверительные интервалы обозначены вертикальными прямоугольниками. Второй тип графиков позволяет визуально сравнить методы при разбиении данных на тренировочную и контрольную части в разных пропорциях. По оси абсцисс расположены значения относительного размера обучающей выборки, по оси ординат – значения метрик качества. Для каждого метода на графике изображаются три ломанные линии, построенные по точкам, соответствующих измерениями. Каждому измерению соответствуют три точки с одинаковым значением по оси абсцисс: левая граница доверительного интервала, среднее значение метрики, правая граница доверительного интервала. Область между линиями, соответствующим левой и правой границам доверительного интервала, закрашена соответствующим цветом.

4.4 Реализация сбора репрезентативного набора данных

Реализован процесс сбора данных. Введена абстракция *SourceData*, используемая методами сэмплинга и для сбора социального графа по заданному множеству идентификаторов целевых пользователей. Абстракция подразумевает следующие операции: *has_attribute(id)* позволяет проверить для заданной вершины, подходит ли она для рассмотрения, известны ли для неё необходимые значения атрибутов; *random_id()* позволяет получить идентификатор случайной вершины, подходящей для рассмотрения; *get_neighs(id)* позволяет получить список идентификаторов соседних вершин. Реализован класс *VkMongoSourceData*, реализующий *SourceData* для дампа социальной сети Вконтакте, хранящейся в СУБД MongoDB. Реализовано два метода сэмплинга:

- RandomWalk – случайные блуждания по графу, переходы от вершины к соседней;
- ForestFire – алгоритм «лесного пожара»;

4.5 Реализация анализа свойств данных

В системе реализован анализ данных с целью оценки свойств гомофилии (H), зависимостей между размером общим контекстом и значениями атрибутов (C), зависимости между специфичностью общего контекста и значениями атрибутов (CS) согласно процессу, описанному в подразделе 2.4.2. Подсчет значений h_i, c_i и cs_i реализован на языке C++, построение графиков реализовано на языке Python 3. На графике изображаются точки, соответствующие значениям $1 - \alpha_i^{\xi}$ и r_i^{ξ}. Смысл этих значений описан в подразделе 2.4.2.

4.6 Реализация веб-сервера для ручного сбора референсных значений атрибутов

В рамках программной системы был реализован веб-сервер, приложение, которое позволяет пользователем указывать значения атрибутов для себя и своих знакомых. Приложение и способы его использования были описаны в подразделе 2.3.2. При реализации были описаны шаблоны HTML страниц. Обработка запросов и генерация HTML страниц из шаблонов реализовано с использованием библиотеки flask [104].

4.7 Используемые библиотеки и программы

При реализации фреймворка и использования его для экспериментального сравнения методов использованы следующие сторонние программы и библиотеки:

- Стандартная библиотека языка Python 3;
- Программа DeepWalk [81] для вычисления статических представлений вершин графа;
- Библиотека машинного обучения sklearn [47] для Python;
- Библиотека алгоритмов бустинга XGBoost [101];

– Библиотека Matplotlib [105] для отрисовки графиков;
– Библиотека DGL [91], фреймворк для построения решений на основе графовых нейронных сетей;
– Библиотека flask [104] для реализации веб-сервера;
– Библиотека для парсинга JSON jute[1] для C++.

4.8 Выводы

В главе описана программная система для предсказания значений демографических атрибутов пользователей социальных сетей по социальному графу. Система позволяет сравнить качество различных методов предсказания значений демографических атрибут и оформить результаты в виде графиков. Кроме того, в системе реализован анализ свойств гомофилии и зависимостей между свойствами общего контекста (размером и специфичностью) и значениями атрибутов. Имеется возможность расширения системы новыми методами, признаками, сэмплинга социальных графов, метриками качества.

Разработанная программная система для предсказания значений атрибутов пользователей социальных сетей по социальному графу позволила экспериментально подтвердить эффективность предложенных в главе 3 методов и их превосходство над существующими методами по качеству предсказания значений демографических атрибутов.

[1] https://github.com/amir-s/jute

Заключение

Основные результаты работы заключаются в следующем.

1. Разработан подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа;
2. В рамках подхода созданы новые методы предсказания значений демографических атрибутов по социальному графу *LP-CS*, *LP-CS-Gen*, *Distr2-CS+DW[n]-XGB*, *GConv-CS[n]*, *Distr2-CS-XGB*, превосходящие по качеству существующие аналоги; даны рекомендации по их применению;
3. Реализована программная система предсказания значений атрибутов пользователей социальных сетей по социальному графу, позволившая экспериментально подтвердить превосходство созданных методов над существующими аналогами по качеству решения задачи.

В рамках обзора были рассмотрены методы определения значений демографических атрибутов пользователей по текстам сообщений и социальному графу, были выявлены недостатки существующих методов. Было описано и определено свойство специфичности контекста вершины графа для заданного атрибута. На нескольких социальных графах, собранных из реальных социальных сетей, экспериментально показано, что специфичность контекста может быть использована для предсказания значений атрибутов. Предложен подход для предсказания значений атрибутов пользователей на основе специфичности контекста и несколько методов на основе подхода. Методы *LP-CS* и *LP-CS-Gen* являются вариацией алгоритма распространения меток. В методах *Distr2-CS-XGB* и *Distr2-CS+DW[n]-XGB* использовались предложенные признаки *Distr2-CS*, основанные на специфичности контекста Метод *GConv-CS[n]* основан на свёрточных графовых нейронных сетях. Даны рекомендации по применению разработанных методов.

Автор выражает благодарность Турдакову Денису за научное руководство, Дробышевскому Михаилу за активное содействие в получении результатов и развитие идей, а также другим коллегам отдела информационных систем ИСП РАН, принимавшим участие в обсуждениях в процессе работы над диссертацией. Также автор выражает признательность авторам шаблона

Russian-Phd-LaTeX-Dissertation-Template за помощь в оформлении диссертации.

Словарь терминов

Социальные медиа – вид массовой коммуникации посредством сети Интернет.

Социальная сеть – онлайн-платформа, которая используется для общения, знакомств, создания социальных отношений между людьми, которые имеют схожие интересы или офлайн-связи, а также для развлечения (музыка, фильмы) и работы.

Пользователь социальной сети – владелец аккаунта, имеющий персональную страницу и возможность создавать социальные отношения в социальной сети.

Демографические атрибуты пользователя – пол, возраст, семейное положение, уровень образования, политические и религиозные взгляды, национальность, интересы и другие.

Демографический профиль пользователя – множество значений демографических атрибутов пользователя.

Стационарные демографические атрибуты пользователя – демографические атрибуты, значения которых редко меняются и актуальны на протяжении жизни пользователей, например, пол, год рождения, род деятельности.

Публичный профиль пользователя – множество явно указанных и публично доступных значений демографических атрибутов пользователя.

Референсное значение атрибута пользователя – значение атрибута пользователя, считающееся истинным.

Сообщество – публичная страница социальной сети, не являющаяся персональной страницей пользователя.

Социальные связи – отношения между страницами социальной сети: дружба между пользователями, подписки пользователей на сообщества.

Социальный граф – ненаправленный граф, моделирующий часть социальной сети, состоящий из вершин, представляющих страницы пользователей, сообществ, организаций и т.д., и рёбер, представляющих социальные связи.

Вершина-пользователь – вершина социального графа, соответствующая персональной странице пользователя.

Первая окрестность вершины графа – множество вершин, соединённых с заданной вершиной ребром.

Соседние вершины, или **соседи** вершины – то же, что первая окрестность.

Двухшаговая окрестность вершины – множество вершин, достижимых из заданной ровно в два шага (перехода по рёбрам), за исключением заданной вершины.

Метка вершины – значение атрибута пользователя, страница которого соответствует заданной вершине (в контексте задачи предсказания значений демографических атрибутов).

Множество **размеченных** вершин – множество вершин с известными метками.

Специфичность контекста вершины – величина, показывающая насколько распределение значений заданного атрибута серди соседей отличается от распределения по всем размеченным вершинам, определяется формулой (2.2).

Специфичность общего контекста пары вершин – величина, представляющая сумму специфичности контекста по всем общим соседям заданных вершин, определена в формуле (2.3).

Классификатор – метод машинного обучения с учителем для решения задач классификации

Регрессор – метод машинного обучения с учителем для решения задач регрессии

n-грамма – последовательность из n подряд идущих слов, встречающаяся в тексте.

Статические векторные представления вершин социального графа – представления вершин графа в виде плотных векторов, извлекаемые только из структуры графа, сохраняющие структурную близость между вершинами.

Свёертка в графовой нейронной сети – преобразование в рамках графовой нейронной сети, вычисляющее представление вершины графа на основе представлений соседей, точное определение дано в формуле (3.6).

Сэмплинг – процесс формирования репрезентативной выборки социальной сети в виде социального графа.

Список литературы

1. *Gomzin A.*, *Drobyshevskiy M.*, *Turdakov D.* Context specificity matters: profile attributes prediction for social network users // Conference on Information Sciences and Systems (CISS), Johns Hopkins University, Mar. 2021. — 2021.
2. *Гомзин А.*, *Кузнецов С.* Методы построения социо-демографических профилей пользователей сети Интернет // Труды Института системного программирования РАН. — 2015. — Т. 27, № 4.
3. *Гомзин А.*, *Кузнецов С.* Метод автоматического определения возраста пользователей с помощью социальных связей // Труды Института системного программирования РАН. — 2016. — Т. 28, № 6.
4. Detection of author's educational level and age based on comments analysis / A. Gomzin [и др.] // Dialogue. — 2018.
5. *Гомзин А. Г.* Предсказание рода деятельности пользователей социальной сети // Ломоносовские чтения-2020. Секция «Вычислительной математики и кибернетики». — М. : М., 2020. — С. 56—57. — (Секция Вычислительной математики и кибернетики).
6. Система сбора пользовательских данных из онлайновых социальных сетей // Свидетельство №2015616047 о государственной регистрации программы для ЭВМ / А. Гомзин [и др.]. — 2015.
7. *Гомзин А.*, *Турдаков Д.*, *др.* Talisman // Свидетельство №2018615539 о государственной регистрации программы для ЭВМ. — 2018.
8. *Гомзин А.*, *Турдаков Д.* Веб-приложение для разметки рода деятельности пользователей социальной сети // Свидетельство №2019661808 о государственной регистрации программы для ЭВМ. — 2019.
9. *Гомзин А.*, *Турдаков Д.* Программное средство методов предсказания рода деятельности пользователя социальной сети по его социальным связям // Свидетельство №2019663796 о государственной регистрации программы для ЭВМ. — 2019.

10. *Гомзин А.*, *Дробышевский М.*, *Турдаков Д.* Фреймворк для сравнения методов предсказания значений атрибутов пользователей социальных сетей // Свидетельство №2020666741 о государственной регистрации программы для ЭВМ. — 2020.

11. *Pennebaker J. W.*, *King L. A.* Linguistic styles: language use as an individual difference. // Journal of personality and social psychology. — 1999. — Т. 77, № 6. — С. 1296.

12. *Goldberg L. R.* The development of markers for the Big-Five factor structure. // Psychological assessment. — 1992. — Т. 4, № 1. — С. 26.

13. Gender, genre, and writing style in formal written texts / S. Argamon [и др.] // TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-. — 2003. — Т. 23, № 3. — С. 321—346.

14. Gender differences in language use: An analysis of 14,000 text samples / M. L. Newman [и др.] // Discourse Processes. — 2008. — Т. 45, № 3. — С. 211—236.

15. Language and gender author cohort analysis of e-mail for computer forensics / O. Y. de Vel [и др.]. — 2002.

16. *Herring S. C.*, *Paolillo J. C.* Gender and genre variation in weblogs // Journal of Sociolinguistics. — 2006. — Т. 10, № 4. — С. 439—459.

17. *Burger J. D.*, *Henderson J. C.* An Exploration of Observable Features Related to Blogger Age. // AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. — 2006. — С. 15—20.

18. Effects of Age and Gender on Blogging. / J. Schler [и др.] // AAAI spring symposium: Computational approaches to analyzing weblogs. Т. 6. — 2006. — С. 199—205.

19. *Mesterharm C.* A multi-class linear learning algorithm related to winnow // Advances in Neural Information Processing Systems. — 2000. — С. 519—525.

20. *Yan X.*, *Yan L.* Gender Classification of Weblog Authors // AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. — 2006. — С. 228—230.

21. *Nowson S.*, *Oberlander J.* The Identity of Bloggers: Openness and Gender in Personal Weblogs. // AAAI spring symposium: Computational approaches to analyzing weblogs. — 2006. — С. 163—167.

22. *Wilson M.* MRC psycholinguistic database: Machine-usable dictionary, version 2.00 // Behavior research methods, instruments, & computers. — 1988. — Т. 20, № 1. — С. 6—10.

23. *Cheng N.*, *Chandramouli R.*, *Subbalakshmi K.* Author gender identification from text // Digital Investigation. — 2011. — Т. 8, № 1. — С. 78—88.

24. *Nguyen D.*, *Smith N. A.*, *Rosé C. P.* Author age prediction from text using linear regression // Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. — Association for Computational Linguistics. 2011. — С. 115—123.

25. Feature-rich part-of-speech tagging with a cyclic dependency network / K. Toutanova [и др.] // Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1. — Association for Computational Linguistics. 2003. — С. 173—180.

26. Discriminating gender on Twitter / J. D. Burger [и др.] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics. 2011. — С. 1301—1309.

27. *Littlestone N.* Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm // Machine learning. — 1988. — Т. 2, № 4. — С. 285—318.

28. Predicting the political alignment of twitter users / M. D. Conover [и др.] // Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. — IEEE. 2011. — С. 192—199.

29. *Raghavan U. N.*, *Albert R.*, *Kumara S.* Near linear time algorithm to detect community structures in large-scale networks // Physical review E. — 2007. — Т. 76, № 3. — С. 036106.

30. Classifying latent user attributes in twitter / D. Rao [и др.] // Proceedings of the 2nd international workshop on Search and mining user-generated contents. — ACM. 2010. — С. 37—44.

31. *Peersman C.*, *Daelemans W.*, *Van Vaerenbergh L.* Predicting age and gender in online social networks // Proceedings of the 3rd international workshop on Search and mining user-generated contents. — ACM. 2011. — С. 37—44.

32. *Manning C. D.* Sch ütze, H.(2000). Foundations of statistical natural language processing. — 2001.

33. Gender identification on twitter using the modified balanced winnow / W. Deitrick [и др.]. — 2012.

34. *Miller Z.*, *Dickinson B.*, *Hu W.* Gender prediction on twitter using stream algorithms with n-gram character features. — 2012.

35. *Preoţiuc-Pietro D.*, *Lampos V.*, *Aletras N.* An analysis of the user occupational class through Twitter content // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — 2015. — С. 1754—1764.

36. *Bouma G.* Normalized (pointwise) mutual information in collocation extraction // Proceedings of GSCL. — 2009. — С. 31—40.

37. *Ng A. Y.*, *Jordan M. I.*, *Weiss Y.* On spectral clustering: Analysis and an algorithm // Advances in neural information processing systems. — 2002. — С. 849—856.

38. Distributed representations of words and phrases and their compositionality / T. Mikolov [и др.] // Advances in neural information processing systems. — 2013. — С. 3111—3119.

39. On the use of URLs and hashtags in age prediction of Twitter users / A. Pandya [и др.] // 2018 IEEE International Conference on Information Reuse and Integration (IRI). — IEEE. 2018. — С. 62—69.

40. *Коршунов А.*, *Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. — 2012. — Т. 23.

41. *Воронцов К.* Вероятностное тематическое моделирование // Москва. — 2013.

42. *Воронцов К.*, *Потапенко А.* Аддитивная регуляризация тематических моделей // Доклады Академии наук. Т. 456. — 2014. — С. 268—271.

43. *Uteuov A.* Topic model for online communities' interests prediction // Procedia Computer Science. — 2019. — Т. 156. — С. 204—213.

44. *Смелик Н. Д., Фильченков А. А.* Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления // Машинное обучение и анализ данных. — 2016. — Т. 2, № 4. — С. 421—441.

45. *Ljubešić N., Fišer D.* Private or corporate? Predicting user types on Twitter // Proceedings of the 2nd workshop on noisy user-generated text (WNUT). — 2016. — С. 4—12.

46. *Loper E., Bird S.* NLTK: the natural language toolkit // arXiv preprint cs/0205028. — 2002.

47. Scikit-learn: Machine Learning in Python / F. Pedregosa [и др.] // Journal of Machine Learning Research. — 2011. — Т. 12. — С. 2825—2830.

48. Pytorch: An imperative style, high-performance deep learning library / A. Paszke [и др.] // Advances in neural information processing systems. — 2019. — С. 8026—8037.

49. Keras / F. Chollet [и др.]. — 2015. — URL: https://github.com/fchollet/keras.

50. Enriching Word Vectors with Subword Information / P. Bojanowski [и др.] // arXiv preprint arXiv:1607.04606. — 2016.

51. *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural computation. — 1997. — Т. 9, № 8. — С. 1735—1780.

52. An algorithm for suffix stripping. / M. F. Porter [и др.] // Program. — 1980. — Т. 14, № 3. — С. 130—137.

53. *Губанов Д. А., Чхартишвили А. Г.* Связи дружбы и комментирования пользователей социальной сети Facebook // Управление большими системами: сборник трудов. — 2014. — № 52.

54. *Girvan M., Newman M. E.* Community structure in social and biological networks // Proceedings of the national academy of sciences. — 2002. — Т. 99, № 12. — С. 7821—7826.

55. *Zhu X., Ghahramani Z.* Learning from labeled and unlabeled data with label propagation. — 2002.

56. *Hamilton W. L.*, *Ying R.*, *Leskovec J.* Representation learning on graphs: Methods and applications // arXiv preprint arXiv:1709.05584. — 2017.

57. *Goyal P.*, *Ferrara E.* Graph embedding techniques, applications, and performance: A survey // Knowledge-Based Systems. — 2018. — Т. 151. — С. 78—94.

58. Graph neural networks: A review of methods and applications / J. Zhou [и др.] // arXiv preprint arXiv:1812.08434. — 2018.

59. *Kipf T. N.*, *Welling M.* Semi-supervised classification with graph convolutional networks // arXiv preprint arXiv:1609.02907. — 2016.

60. Pregel: a system for large-scale graph processing / G. Malewicz [и др.] // Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. — 2010. — С. 135—146.

61. *Jurgens D.* That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. // ICWSM. — 2013. — Т. 13, № 13. — С. 273—282.

62. A study of age gaps between online friends / L. Liao [и др.] // Proceedings of the 25th ACM conference on Hypertext and social media. — 2014. — С. 98—106.

63. *Bron C.*, *Kerbosch J.* Algorithm 457: finding all cliques of an undirected graph // Communications of the ACM. — 1973. — Т. 16, № 9. — С. 575—577.

64. *Li R.*, *Wang C.*, *Chang K. C.-C.* User profiling in an ego network: co-profiling attributes and relationships // Proceedings of the 23rd international conference on World wide web. — 2014. — С. 819—830.

65. *Dougnon R. Y.*, *Fournier-Viger P.*, *Nkambou R.* Inferring user profiles in online social networks using a partial social graph // Canadian Conference on Artificial Intelligence. — Springer. 2015. — С. 84—99.

66. *Filippova K.* User demographics and language in an implicit social network // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — Association for Computational Linguistics. 2012. — С. 1478—1488.

67. Twitter polarity classification with label propagation over lexical links and the follower graph / M. Speriosu [и др.] // Proceedings of the First workshop on Unsupervised Learning in NLP. — Association for Computational Linguistics. 2011. — С. 53—63.

68. *Talukdar P. P.*, *Crammer K.* New regularized algorithms for transductive learning // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — Springer. 2009. — С. 442—457.

69. You are who you know: inferring user profiles in online social networks / A. Mislove [и др.] // Proceedings of the third ACM international conference on Web search and data mining. — ACM. 2010. — С. 251—260.

70. *Clauset A.*, *Newman M. E.*, *Moore C.* Finding community structure in very large networks // Physical review E. — 2004. — Т. 70, № 6. — С. 066111.

71. Estimating age privacy leakage in online social networks / R. Dey [и др.] // 2012 proceedings ieee infocom. — IEEE. 2012. — С. 2836—2840.

72. *Han J.*, *Wen J.-R.*, *Pei J.* Within-network classification using radius-constrained neighborhood patterns // Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. — 2014. — С. 1539—1548.

73. Finding Organizational Accounts Based on Structural and Behavioral Factors on Twitter / S. Alzahrani [и др.] // International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. — Springer. 2018. — С. 164—175.

74. The PageRank citation ranking: Bringing order to the web.Тех. отч. / L. Page [и др.] ; Stanford InfoLab. — 1999.

75. *Seidman S. B.* Network structure and minimum degree // Social networks. — 1983. — Т. 5, № 3. — С. 269—287.

76. *Watts D. J.*, *Strogatz S. H.* Collective dynamics of 'small-world'networks // nature. — 1998. — Т. 393, № 6684. — С. 440—442.

77. *Kosinski M.*, *Stillwell D.*, *Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Proceedings of the national academy of sciences. — 2013. — Т. 110, № 15. — С. 5802—5805.

78. Идеальный политик для социальной сети: подход к анализу идеологических предпочтений пользователей / Л. Г. Бызов [и др.] // Проблемы управления. — 2020. — Т. 4, № 0. — С. 15—26.

79. A multi-source integration framework for user occupation inference in social media systems / Y. Huang [и др.] // World Wide Web. — 2015. — Т. 18, № 5. — С. 1247—1267.

80. *Newman M. E.* Modularity and community structure in networks // Proceedings of the national academy of sciences. — 2006. — Т. 103, № 23. — С. 8577—8582.

81. *Perozzi B.*, *Al-Rfou R.*, *Skiena S.* DeepWalk: Online Learning of Social Representations // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, New York, USA : ACM, 2014. — С. 701—710. — (KDD '14). — URL: http://doi.acm.org/10.1145/2623330.2623732.

82. *Perozzi B.*, *Skiena S.* Exact age prediction in social networks // Proceedings of the 24th International Conference on World Wide Web. — ACM. 2015. — С. 91—92.

83. Inferring user demographics and social strategies in mobile social networks / Y. Dong [и др.] // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2014. — С. 15—24.

84. *Takac L.*, *Zabovsky M.* Data analysis in public social networks // International Scientific Conference and International Workshop Present Day Trends of Innovations. Т. 1. — 2012.

85. *Aletras N.*, *Chamberlain B. P.* Predicting twitter user socioeconomic attributes with network and language information // Proceedings of the 29th on Hypertext and Social Media. — 2018. — С. 20—24.

86. *Ivanov O. U.*, *Bartunov S. O.* Learning Representations in Directed Networks // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2015. — С. 196—207.

87. *Gutmann M. U.*, *Hyvärinen A.* Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics // Journal of Machine Learning Research. — 2012. — Т. 13, Feb. — С. 307—361.

88. *Трофимович Ю. С.*, *Козлов И. С.*, *Турдаков Д. Ю.* Подходы к определению основного места проживания пользователей социальных сетей на основе социального графа // Труды Института системного программирования РАН. — 2016. — Т. 28, № 6.

89. Demographic Inference on Twitter using Recursive Neural Networks / S. Mac Kim [и др.] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Т. 2. — 2017. — С. 471—477.

90. Learning longer memory in recurrent neural networks / T. Mikolov [и др.] // arXiv preprint arXiv:1412.7753. — 2014.

91. Deep graph library: Towards efficient and scalable deep learning on graphs / M. Wang [и др.] // arXiv preprint arXiv:1909.01315. — 2019.

92. What's in a name: A study of names, gender inference, and gender behavior in facebook / C. Tang [и др.] // International Conference on Database Systems for Advanced Applications. — Springer. 2011. — С. 344—356.

93. *Liu W.*, *Ruths D.* What's in a name? using first names as features for gender inference in twitter // 2013 AAAI Spring Symposium Series. — Citeseer. 2013.

94. *Alowibdi J. S.*, *Buy U. A.*, *Yu P.* Empirical evaluation of profile characteristics for gender classification on twitter // 2013 12th International Conference on Machine Learning and Applications. Т. 1. — IEEE. 2013. — С. 365—369.

95. *McCorriston J.*, *Jurgens D.*, *Ruths D.* Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. // ICWSM. — Citeseer. 2015. — С. 650—653.

96. *Al Zamal F.*, *Liu W.*, *Ruths D.* Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. // ICWSM. — 2012. — Т. 270.

97. Анализ данных (data mining) онлайн социальных сетей с помощью бикластеризации и трикластеризации / Д. Гнатышак [и др.] // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16–20 октября 2012 г., Белгород, Россия). Т. 2. — 2012. — С. 66—73.

98. *Leskovec J.*, *Faloutsos C.* Sampling from large graphs // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2006. — С. 631—636.

99. Walking in facebook: A case study of unbiased sampling of osns / M. Gjoka [и др.] // 2010 Proceedings IEEE Infocom. — Ieee. 2010. — С. 1—9.

100. LINE: Large-scale Information Network Embedding / J. Tang [и др.] // WWW. — ACM. 2015.

101. *Chen T.*, *Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Francisco, California, USA : ACM, 2016. — С. 785—794. — (KDD '16). — URL: http://doi.acm.org/10.1145/2939672.2939785.

102. *Gilyazev R.*, *Turdakov D. Y.* Active Learning and Crowdsourcing: A Survey of Optimization Methods for Data Labeling // Programming and Computer Software. — 2018. — Т. 44, № 6. — С. 476—491.

103. *Student.* The probable error of a mean // Biometrika. — 1908. — С. 1—25.

104. *Grinberg M.* Flask web development: developing web applications with python. — "O'Reilly Media, Inc.", 2018.

105. *Hunter J. D.* Matplotlib: A 2D graphics environment // IEEE Annals of the History of Computing. — 2007. — Т. 9, № 03. — С. 90—95.

Список рисунков

Список таблиц

Приложение А

Экспериментальное сравнение синхронных и асинхронных версий алгоритма распространения меток

В приложении описываются результаты экспериментального сравнения асинхронной и синхронной версии метода распространения меток, описанных в алгоритмах 1 и 2. Для экспериментального сравнения использовались наборы данных twitter (атрибуты род деятельности и доход), vk1 (атриюбуты род деятельности, пол, возраст) и vk2 (атрибуты пол, возраст).

Процесс экспериментального сравнения аналогичен процессу, описанному в разделе 3.4. Сравнивается качество работы асинхронной и синхронной версии алгоритма распространения меток с количеством итераций от 1 до 4.

Результаты экспериментального сравнения представлены на рисунках. Асинхронная версия с n итерациями обозначена как LP-A[n]. Синхронная версия с n итерациями обозначена как LP-S[n]. По результатам сравнения можно сделать вывод, что синхронная версия алгоритма с 2 итерациями показывает наилучшее качество в большинстве случаев.

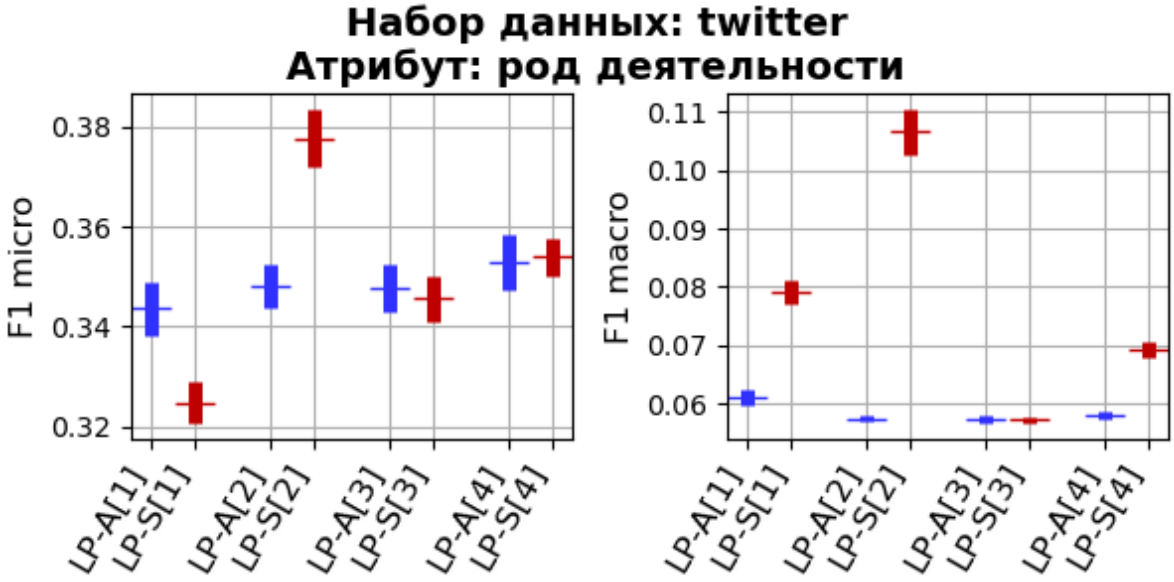

Рисунок А.1 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных twitter; атрибут: род деятельности

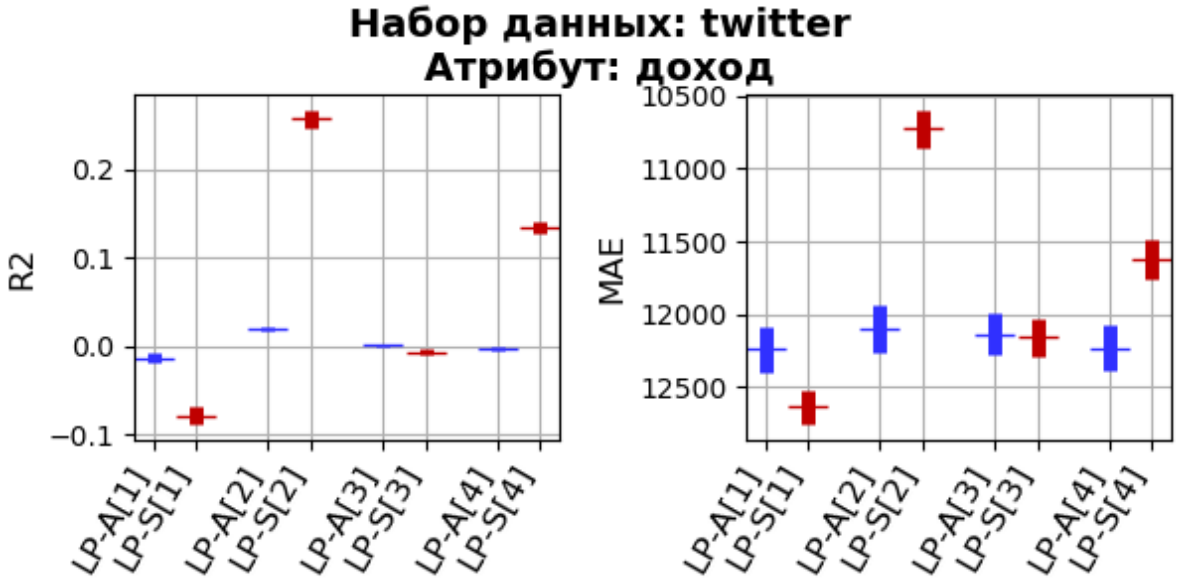

Рисунок А.2 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных twitter; атрибут: доход

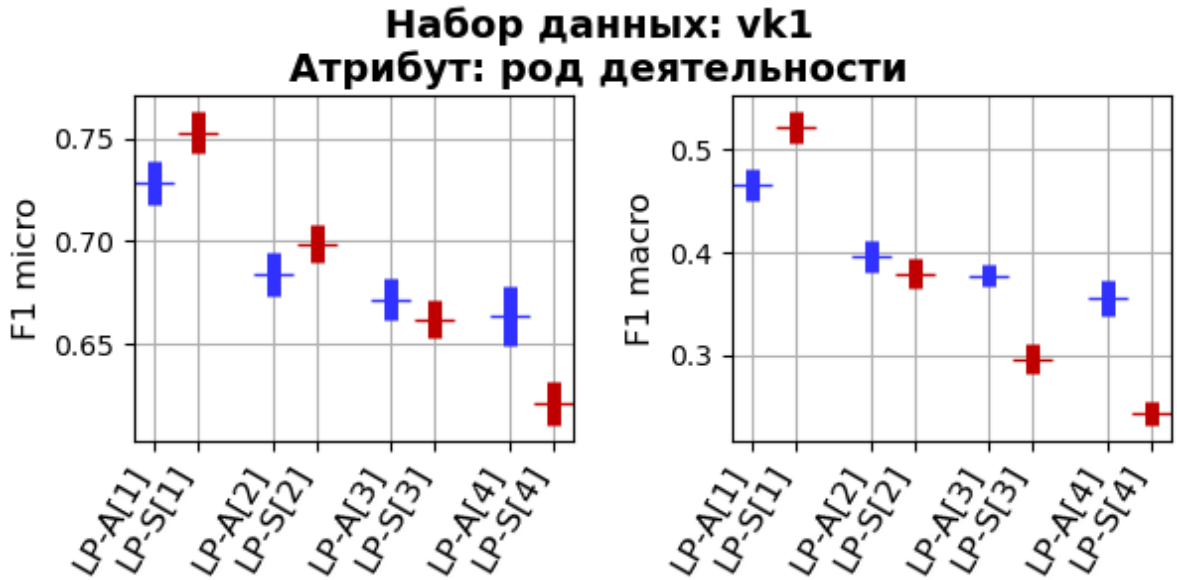

Рисунок А.3 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных vk1; атрибут: род деятельности

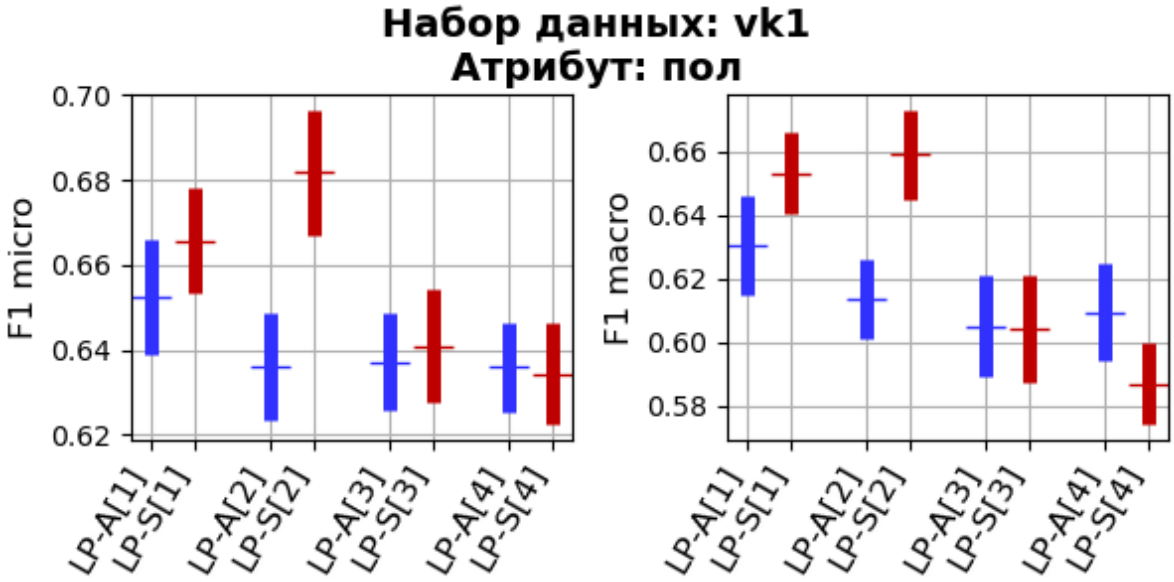

Рисунок А.4 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных vk1; атрибут: пол

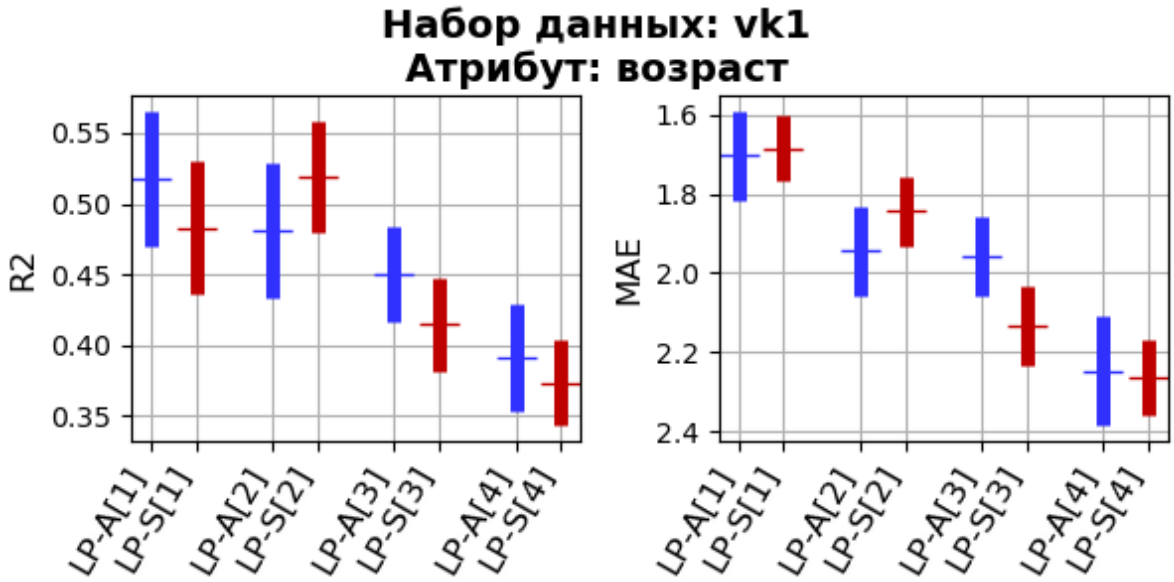

Рисунок А.5 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных vk1; атрибут: возраст

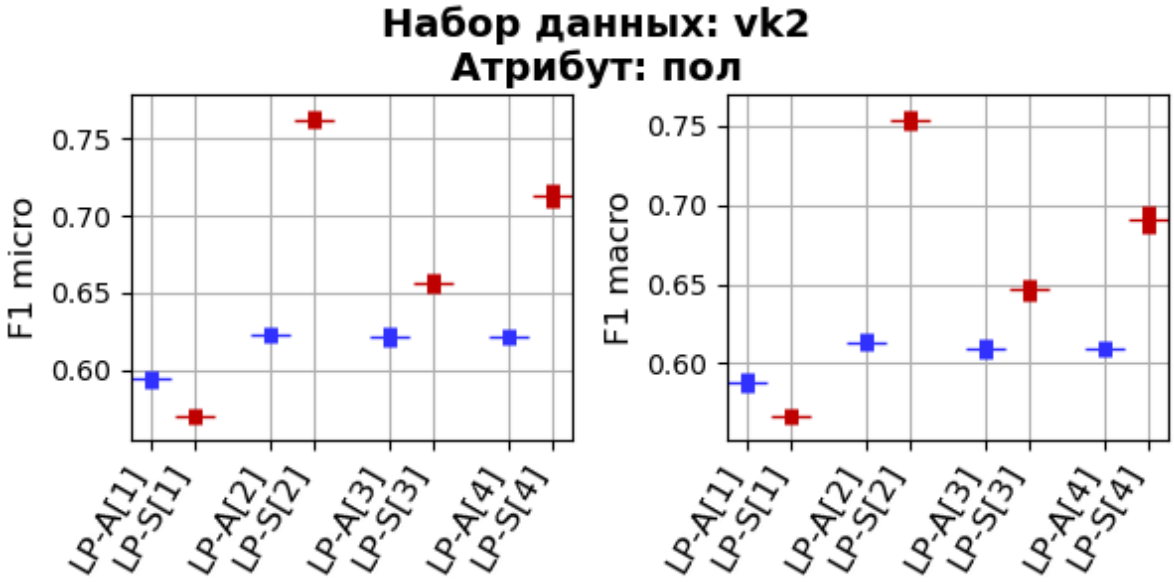

Рисунок А.6 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных vk2; атрибут: пол

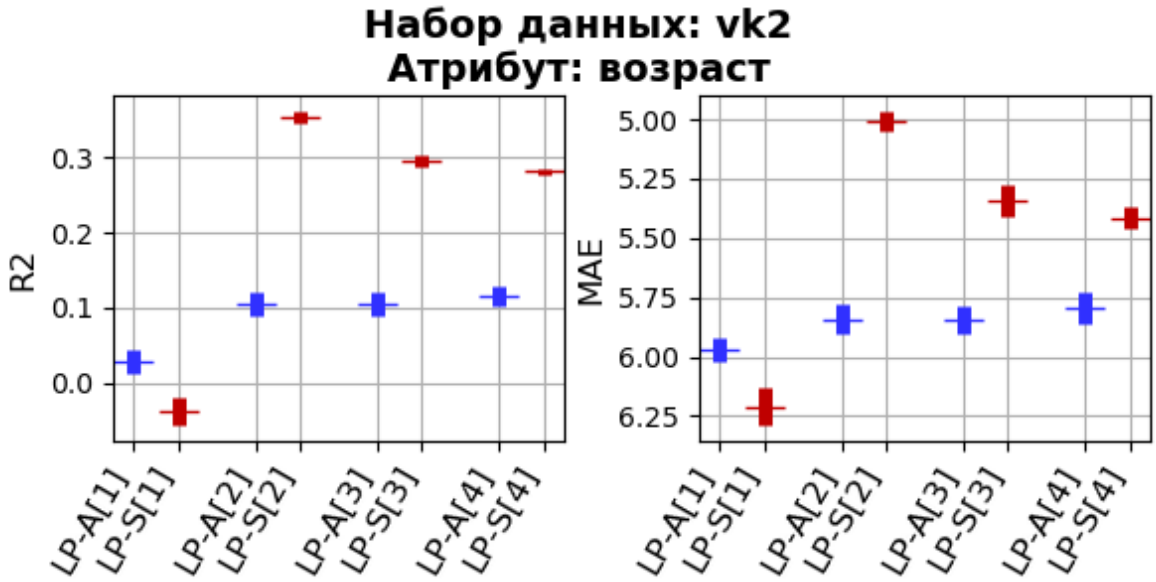

Рисунок А.7 — Качество работы синхронной и асинхронной версии алгоритма распространения меток на наборе данных vk2; атрибут: возраст

Приложение Б

Экспериментальное сравнение методов при различных пропорциях разбиения на тренировочную и тестовую выборки

В приложении описываются результаты экспериментального сравнения методов LP[2], LP-CS[2] (для регрессии), LP-CS-Gen (для классификации), Distr2-CS-XGB, DW[n]-XGB, Distr2-CS+DW[n]-XGB, GConv[n], GConv-CS[n] при различных пропорциях обучающей и тестовой выборок. Для экспериментального сравнения использовались наборы данных twitter (атрибуты род деятельности и доход), vk1 (атриюбуты род деятельности, пол, возраст) и vk2 (атрибуты пол, возраст).

Процесс экспериментального сравнения аналогичен процессу, описанному в разделе 3.4. В качестве доли тренировочных данных использовались следующие значения: $[5\%, 20\%, 35\%, 50\%, 65\%, 80\%, 95\%]$.

Результаты экспериментального сравнения представлены на рисунках. Для каждой пары (набор данных, атрибут) представлено 3 графика. На первом графике сравниваются базовый и модифицированный алгоритмы распространения меток. Второй график показывает качество работы метода, использующего только статические векторные представления вершин, только представления Distr2-CS и их комбинации. Третий график показывает сравнение методов GConv[n] и GConv-CS[n], основанных на графовых нейронных сетях.

По результатам сравнения можно сделать вывод, что в большинстве случаев соотношение качества методов не зависит от доли размеченных данных. Иными словами, независимо от доли размеченных вершин, методы, основанные на специфичности контекста, превосходят по качеству базовые методы, либо показывают качество не хуже, чем базовые методы.

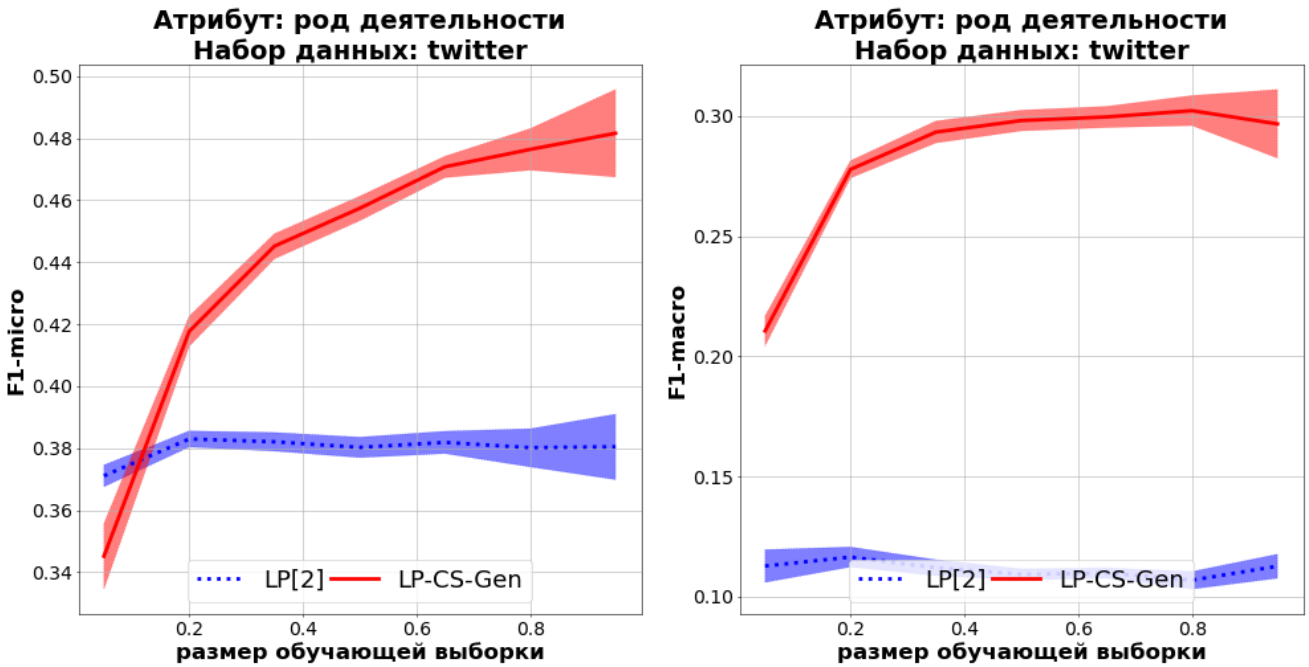

Рисунок Б.1 — Качество предсказания при различных размерах обучающей выборки. Набор данных: twitter, атрибут: род деятельности

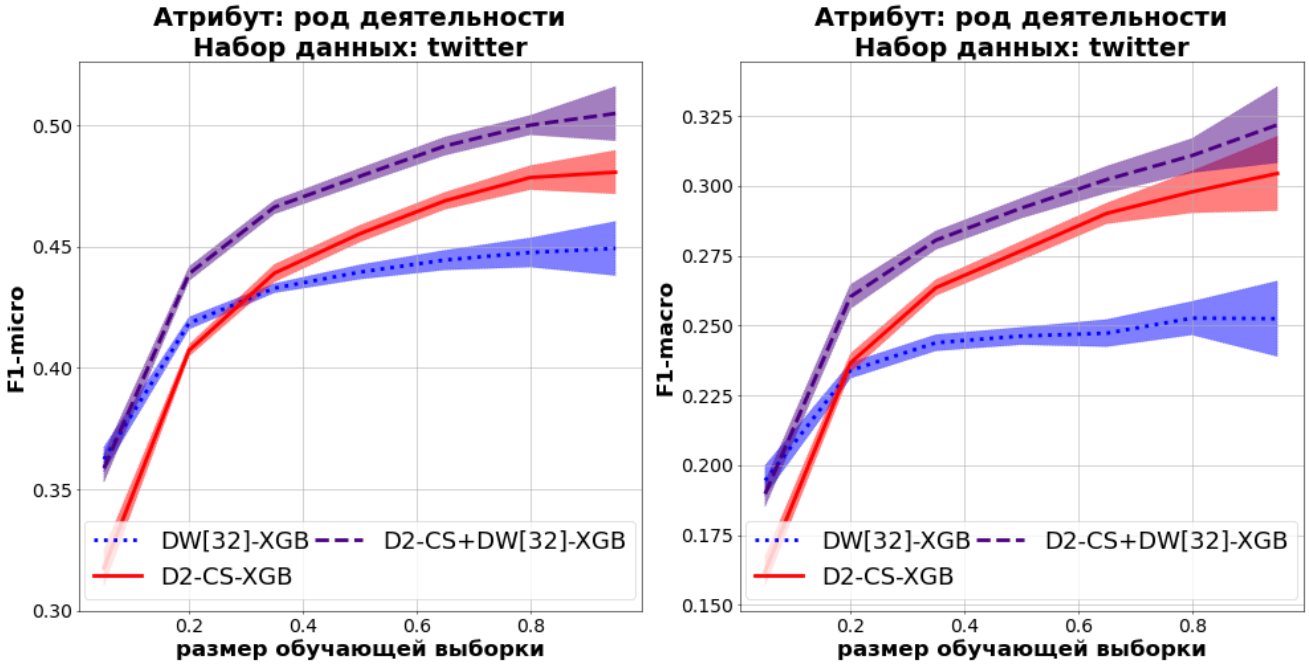

Рисунок Б.2 — Качество предсказания при различных размерах обучающей выборки. Набор данных: twitter, атрибут: род деятельности

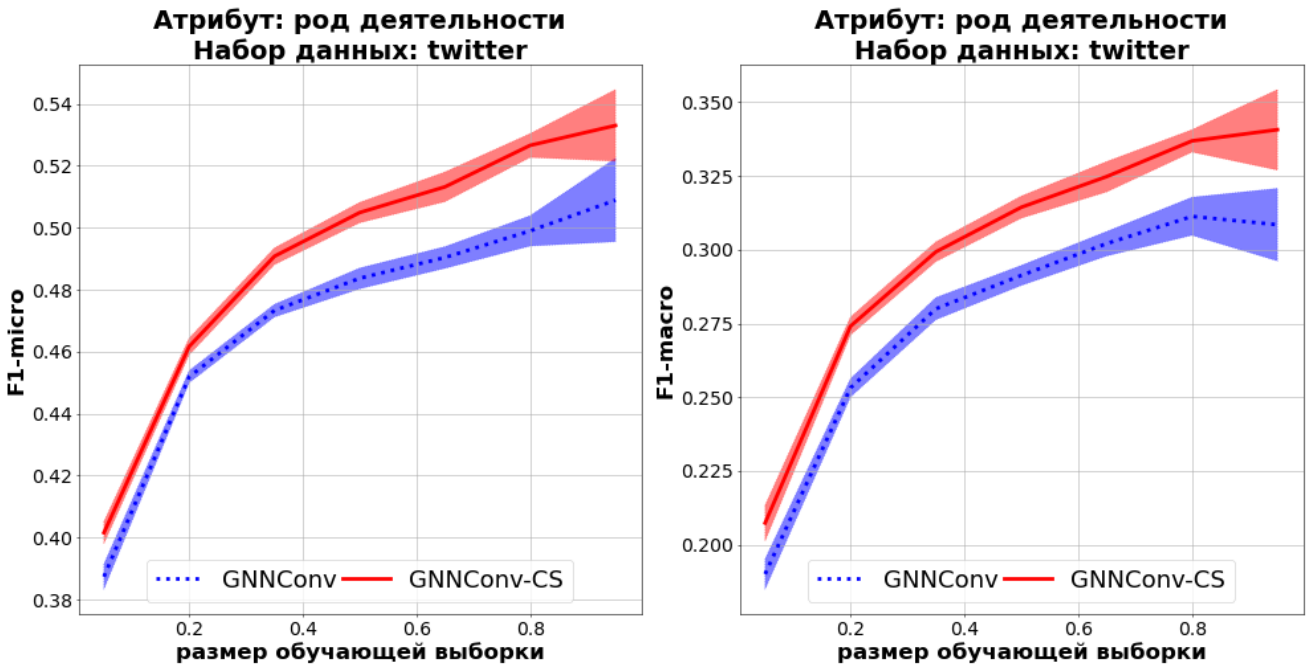

Рисунок Б.3 — Качество предсказания при различных размерах обучающей выборки. Набор данных: twitter, атрибут: род деятельности

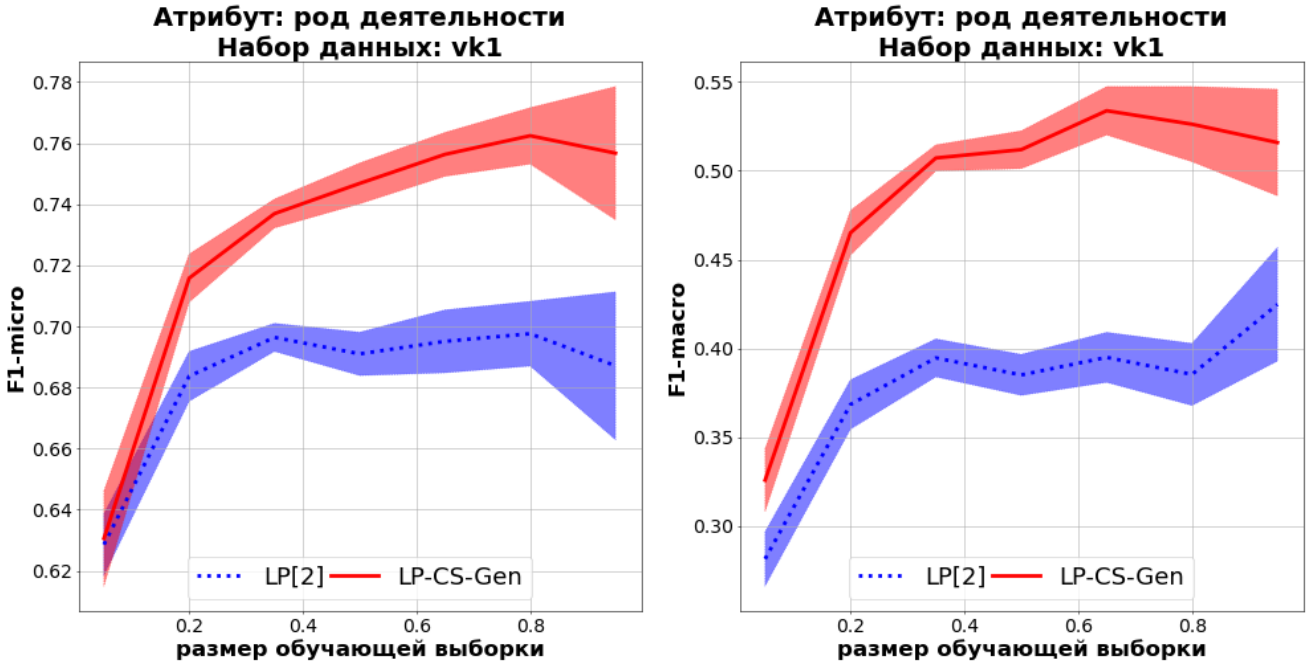

Рисунок Б.4 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: род деятельности

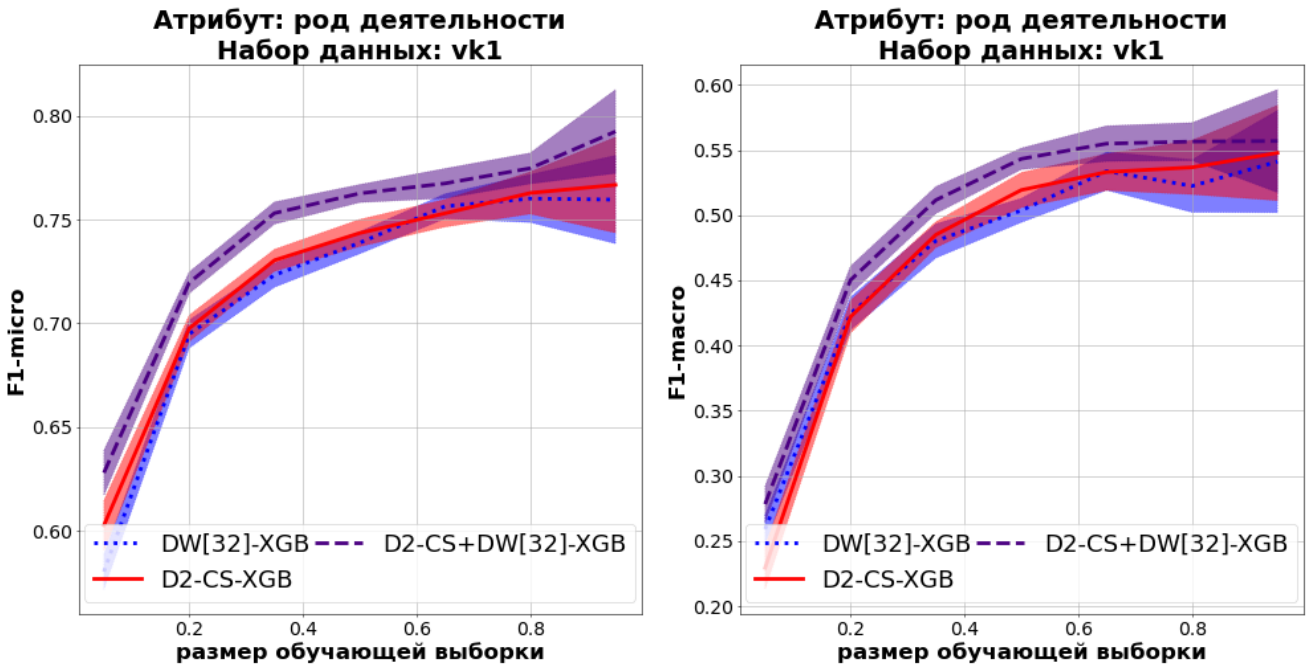

Рисунок Б.5 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: род деятельности

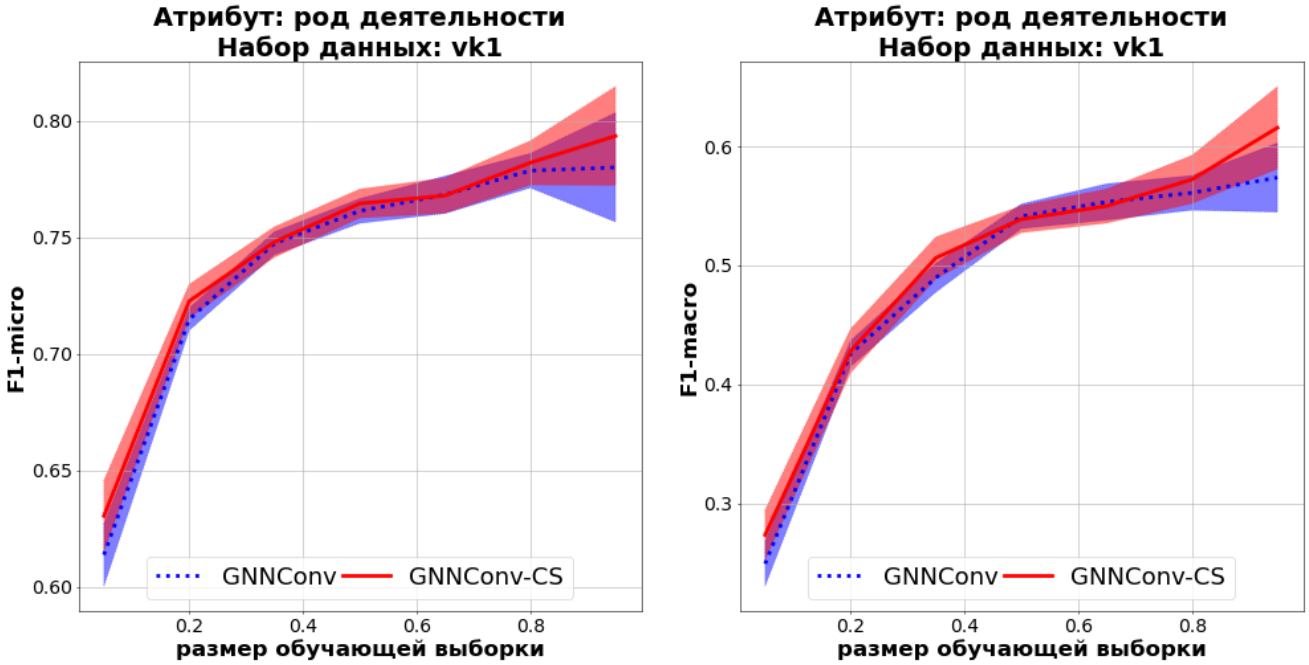

Рисунок Б.6 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: род деятельности

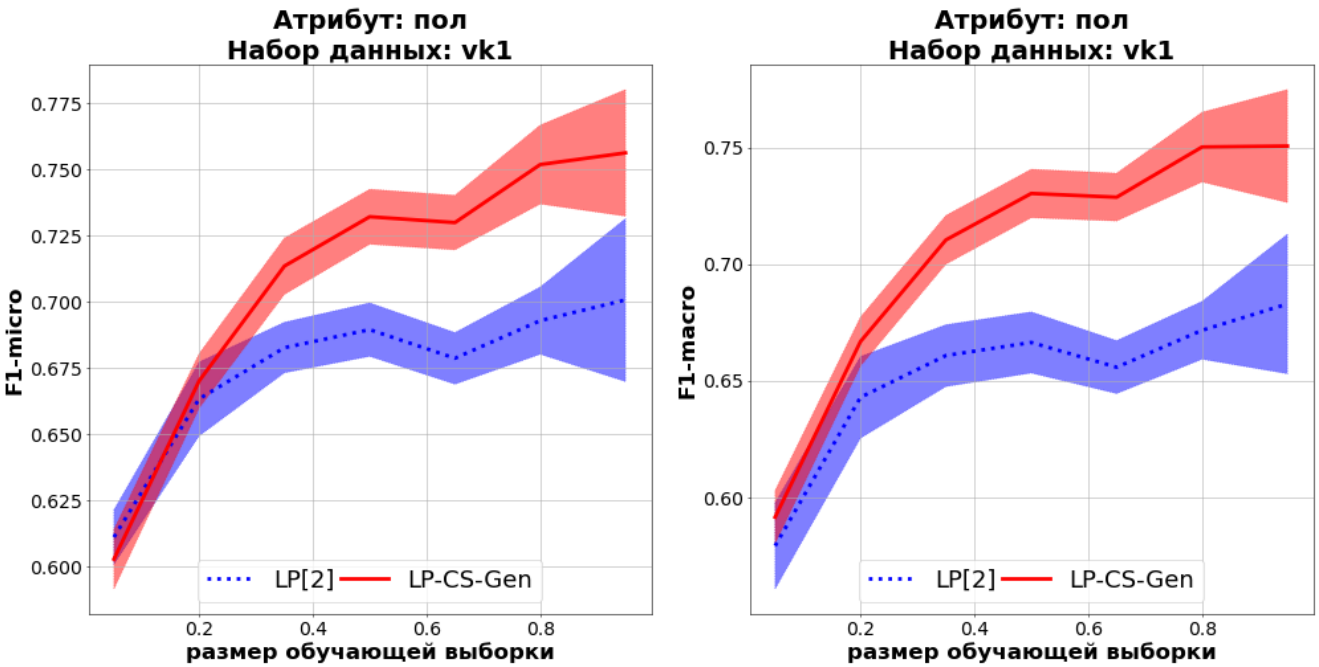

Рисунок Б.7 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: пол

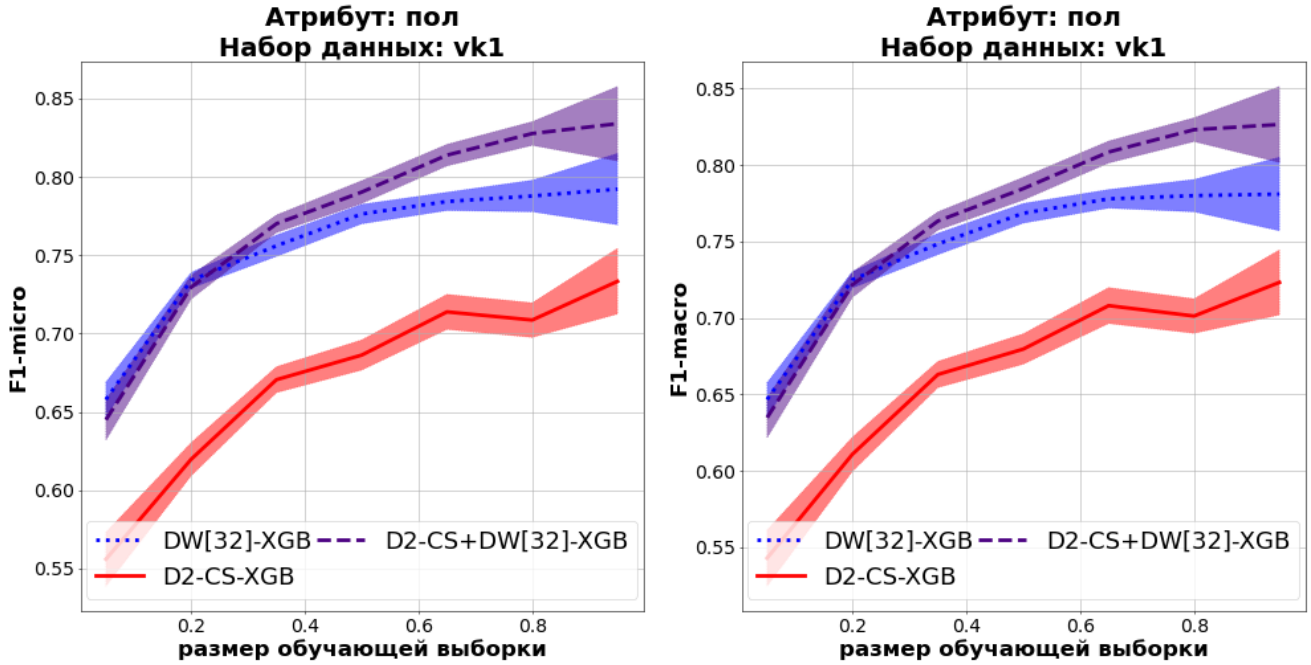

Рисунок Б.8 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: пол

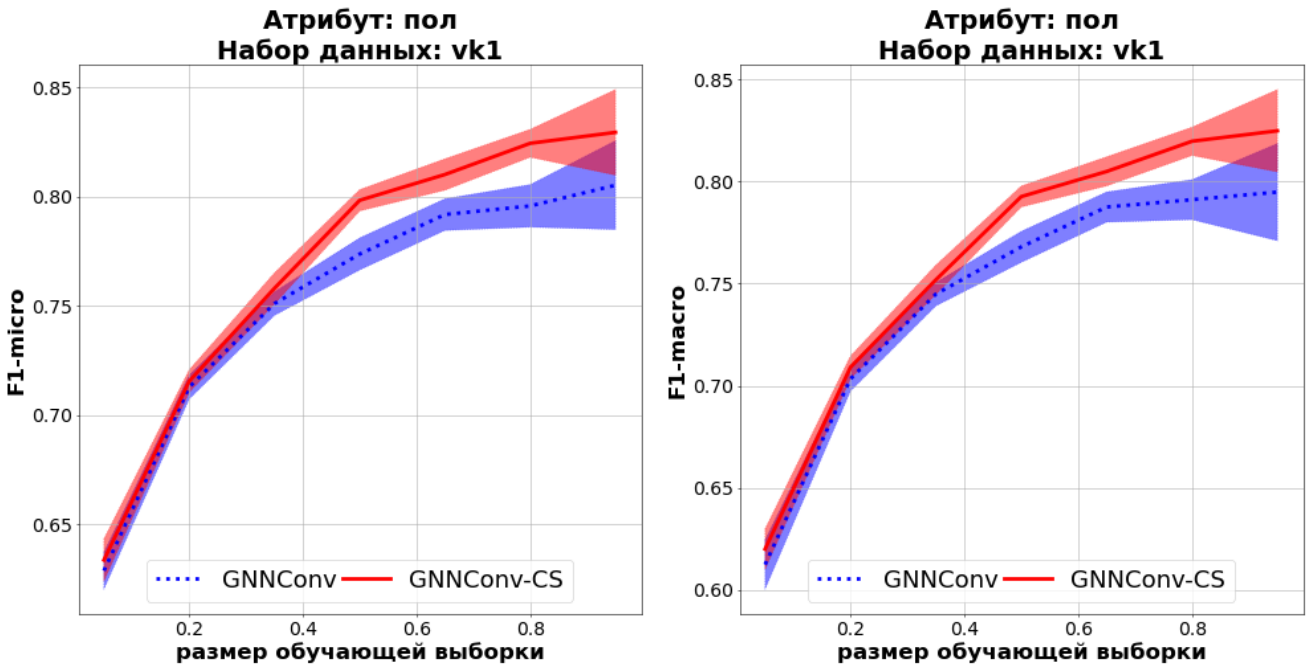

Рисунок Б.9 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: пол

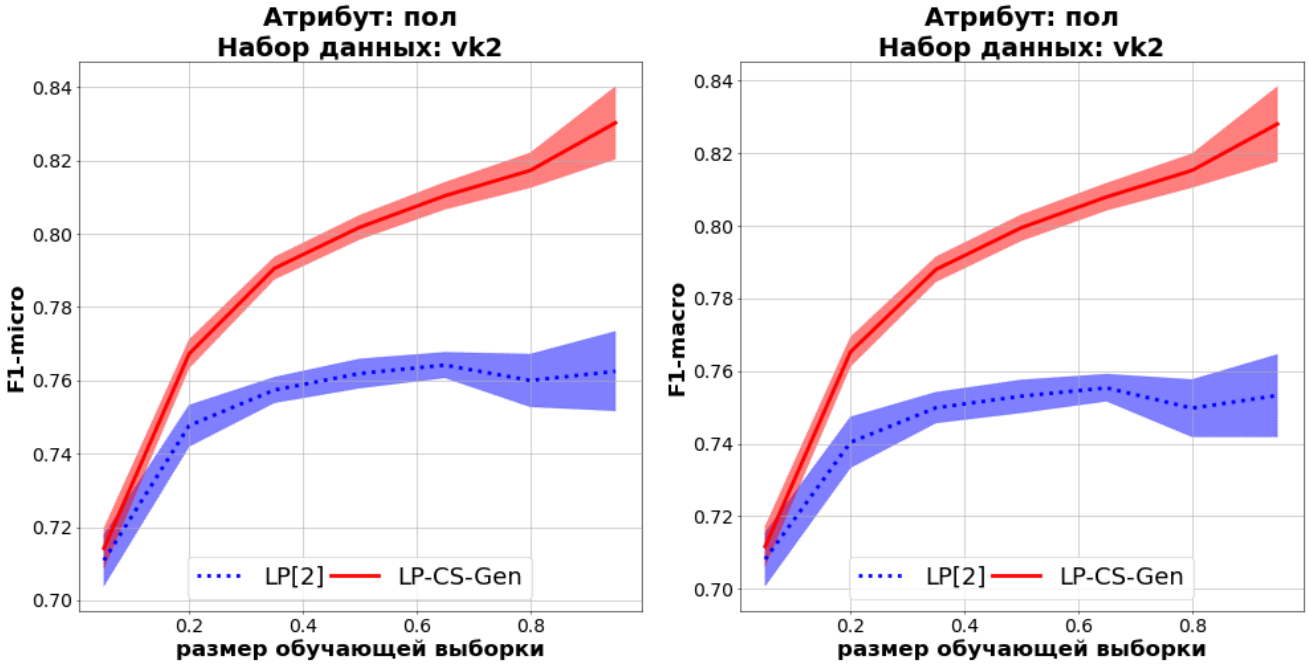

Рисунок Б.10 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk2, атрибут: пол

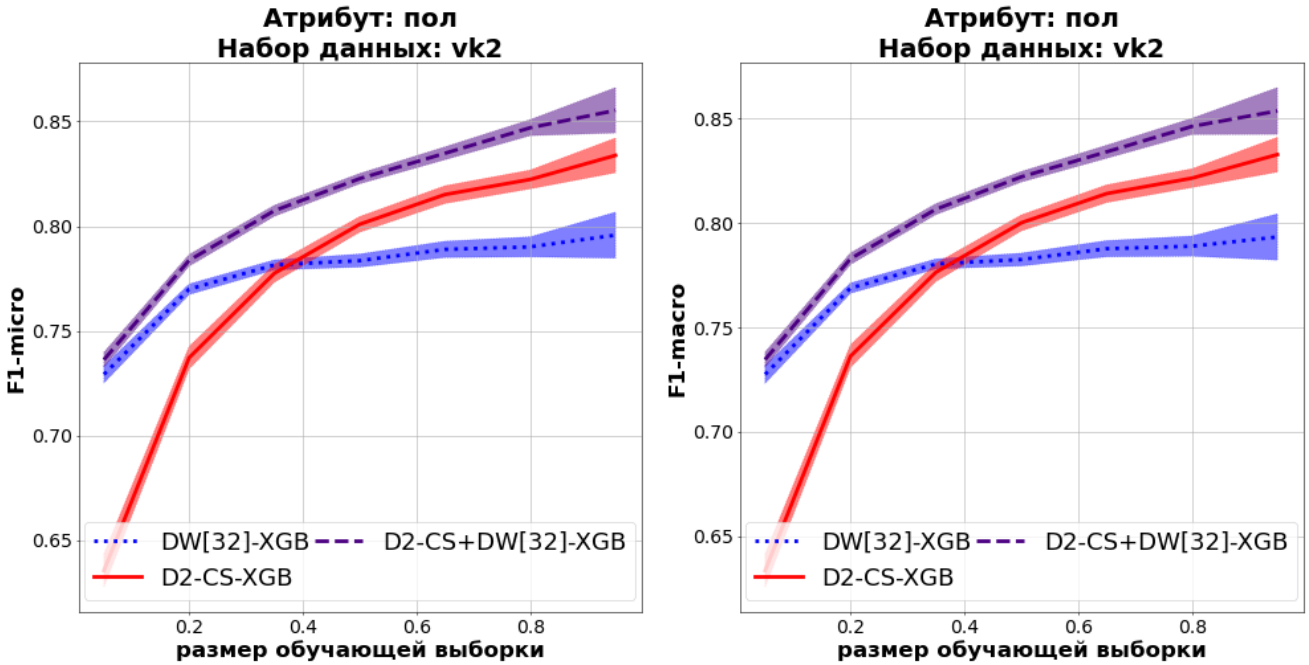

Рисунок Б.11 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk2, атрибут: пол

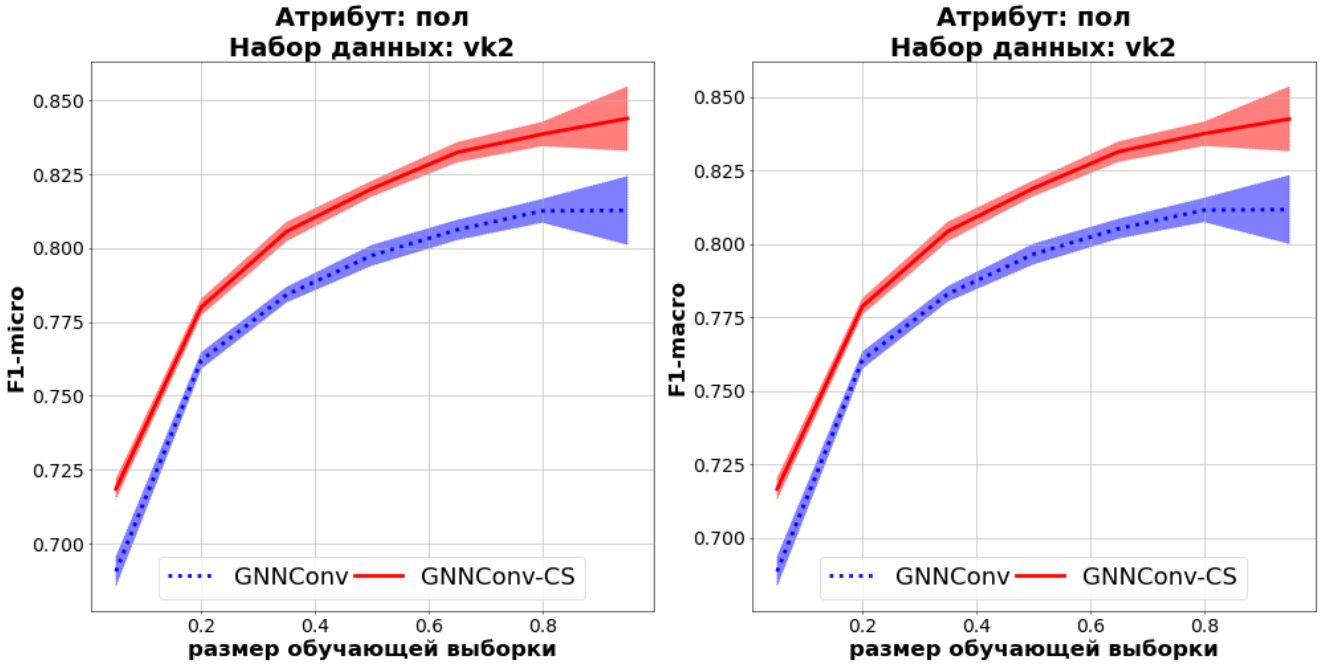

Рисунок Б.12 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk2, атрибут: пол

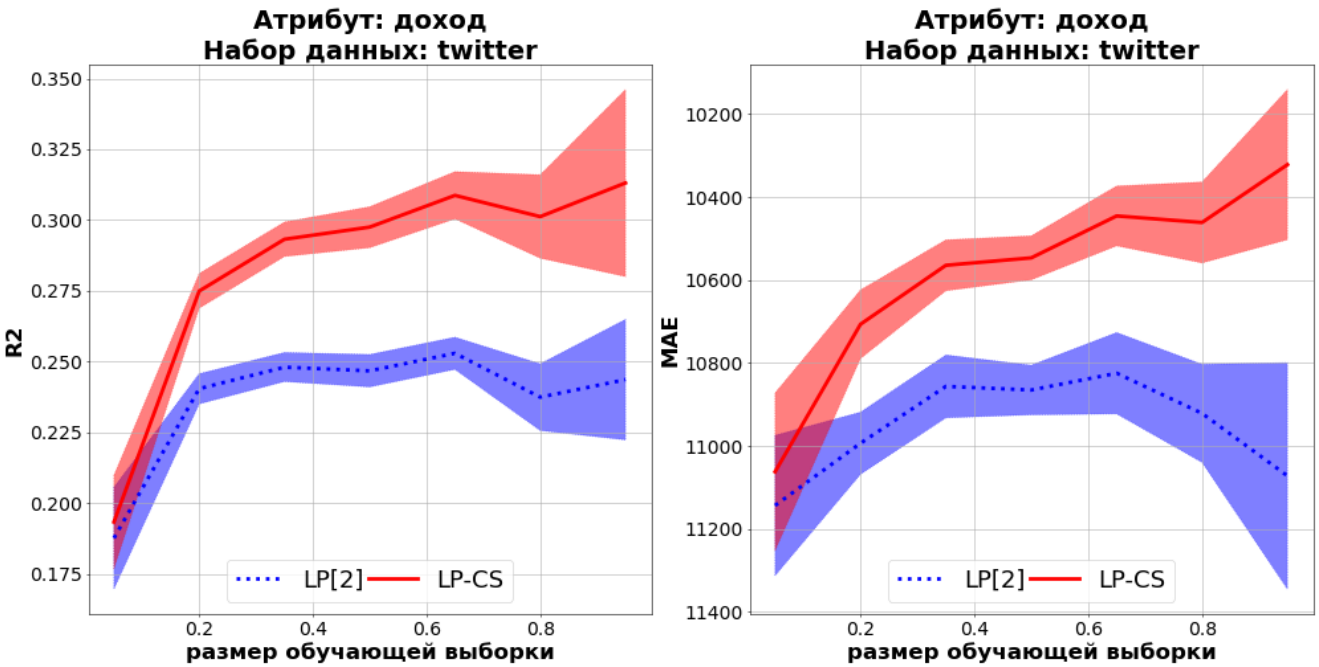

Рисунок Б.13 — Качество предсказания при различных размерах обучающей выборки. Набор данных: twitter, атрибут: доход

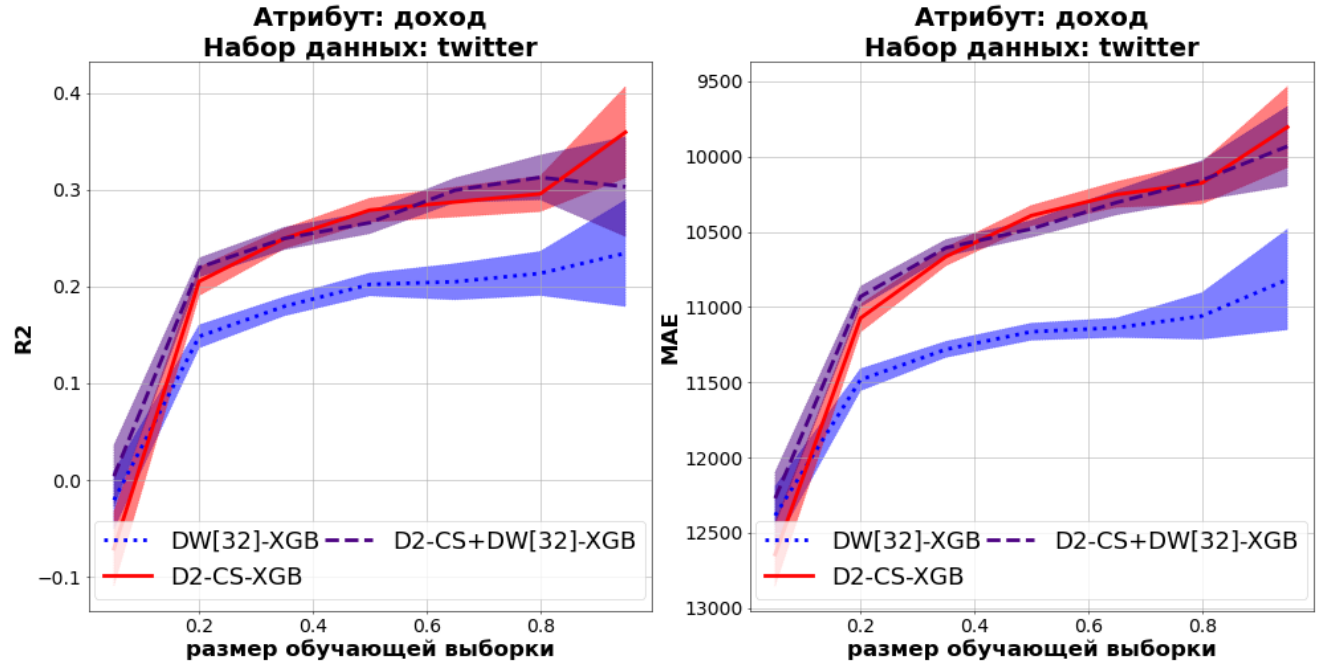

Рисунок Б.14 — Качество предсказания при различных размерах обучающей выборки. Набор данных: twitter, атрибут: доход

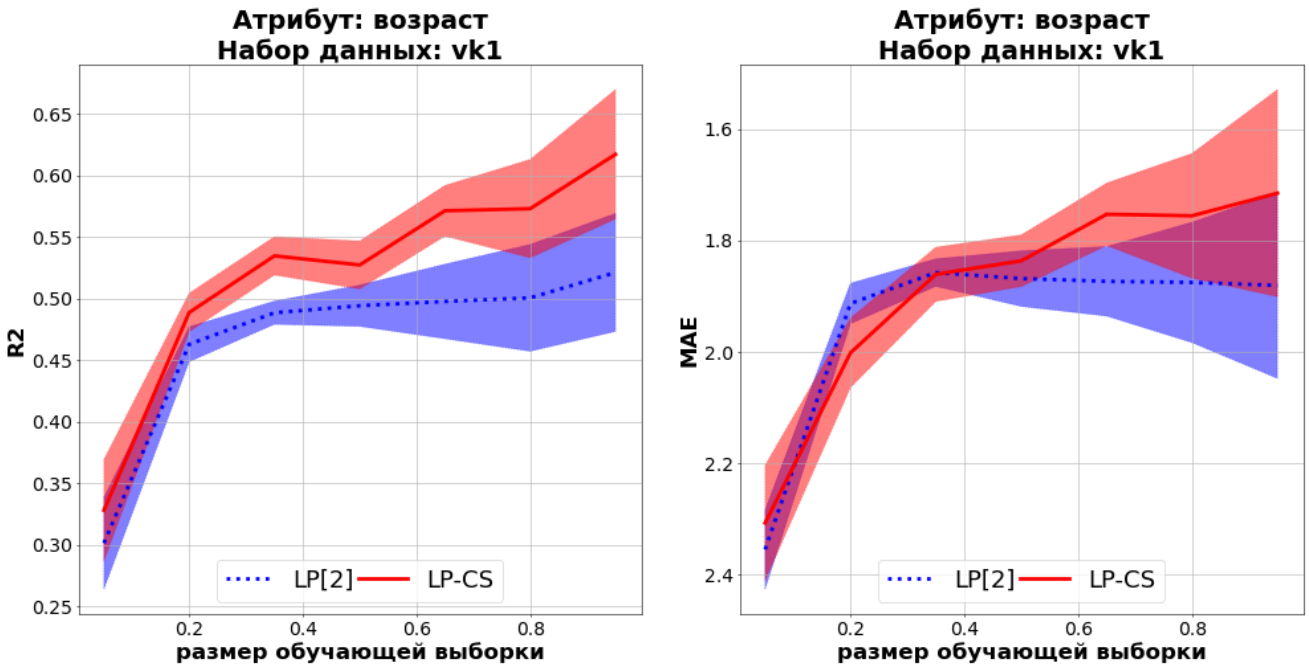

Рисунок Б.15 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: возраст

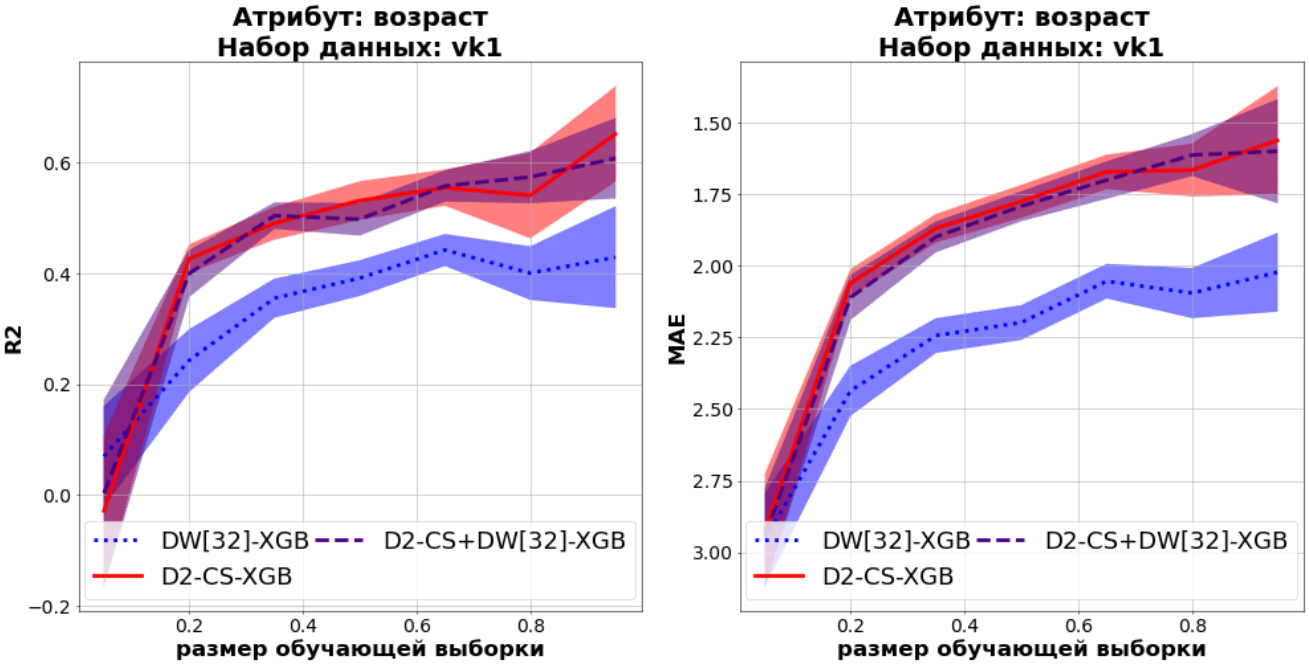

Рисунок Б.16 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk1, атрибут: возраст

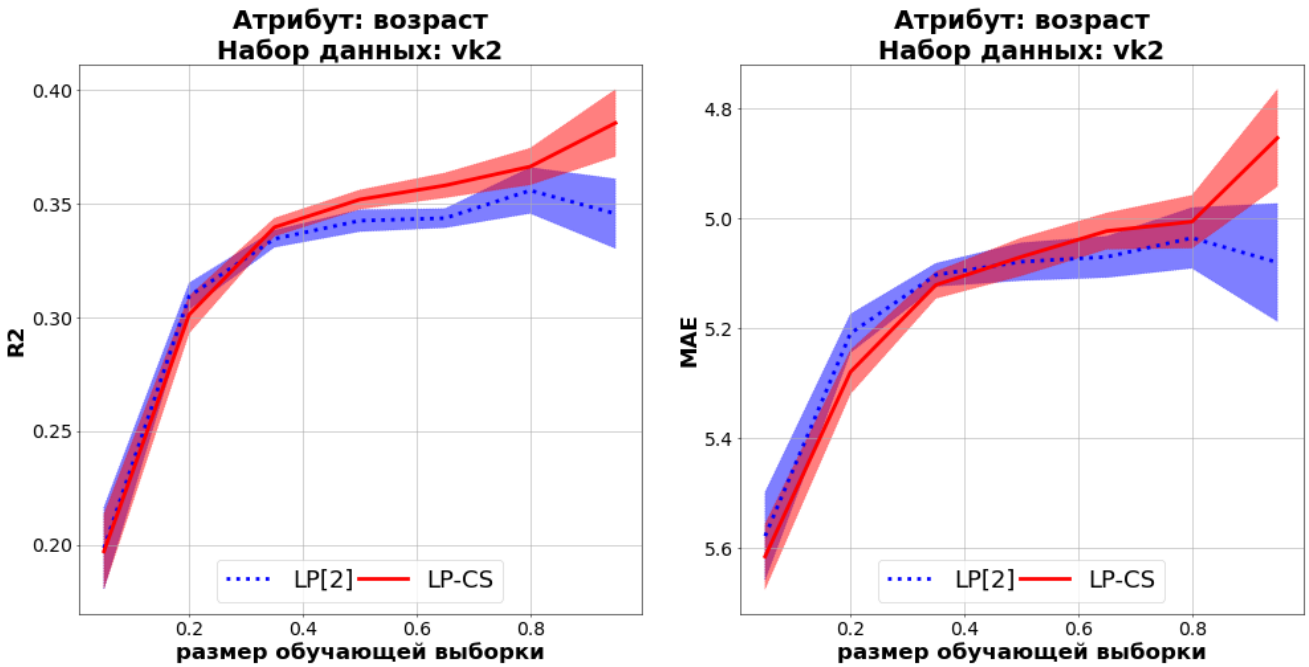

Рисунок Б.17 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk2, атрибут: возраст

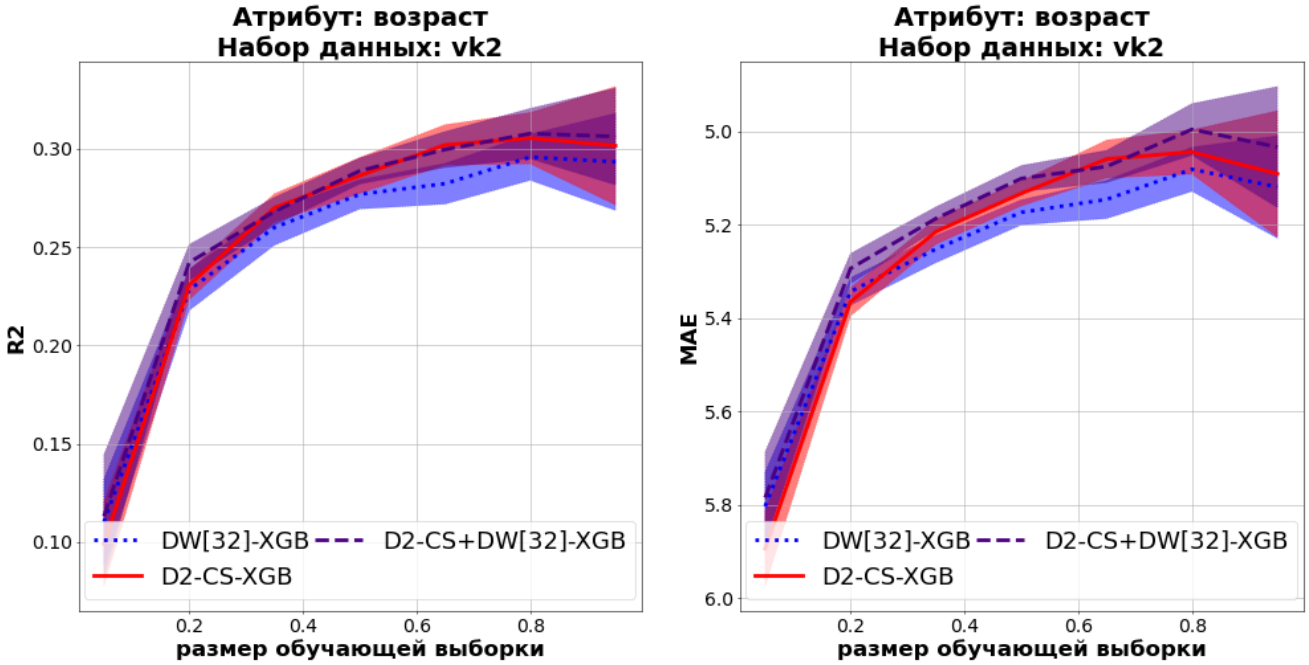

Рисунок Б.18 — Качество предсказания при различных размерах обучающей выборки. Набор данных: vk2, атрибут: возраст