Министерство науки и высшего образования
Российской Федерации
Московский физико-технический институт
(государственный университет)
Факультет управления и прикладной математики
Базовая организация - Институт системного
программирования
им. В.П. Иванникова РАН
Кафедра «Системное программирование»
Специализация «Математические и информационные
технологии (ИСП РАН)»

Ефремова Мария Александровна

Использование распределения подграфов в графе для определения демографических атрибутов пользователей сети Интернет

Выпускная квалификационная работа бакалавра

Научный руководитель:
к.ф.-м.н. Турдаков Денис Юрьевич
Научный консультант:
Дробышевский Михаил Дмитриевич

Москва
2018

Содержание

Аннотация

В сети Интеренет существуют ресурсы, которые позволяют пользователям делиться информацией о себе - заполнять профиль. Эти данные могут оказаться полезными для социологических исследований, а также для работы рекомендательных систем и таргетированной рекламы. Однако некоторые пользователи предпочитают оставлять незаполненными необязательные поля или указывают в профиле неверную информацию, из-за чего возникает задача автоматического предсказания недостающих значений. В данной работе исследуется возможность использования распределений подграфов в графе в качестве признака для описания пользователей в решении задачи автоматического предсказания демографических атрибутов пользователей социальной сети ВКонтакте с помощью методов машинного обучения. Работы, в которых бы уже были использованы распределения подграфов для определения социально-демографических параметров, не найдены, однако есть исследования, которые показывают, что распределения подграфов малых размеров позволяют классифицировать графы по доменам. Это даёт основание прелположить, что распределения подграфов целесообразно применять в описании пользователей интернет-ресурсов. Предложены методы решения данной задачи, основанные на использовании распределений подграфов во второй окрестности пользователя в социальном графе, а именно: формирование признакового описания только на распределениях подграфов, использование конкатенации распределений атрибутов соседей пользователя и распределений подграфов и конкатенации распределений подграфов с центроидами классов. Проводится экспериментальная оценка качества предлагаемых методов. Полученные результаты свидетельствуют о том, что в ходе исследования не удалось выявить зависимость между социально-демографическими характеристиками пользователей и распределениями подграфов в их вторых окрестностях в графе социальных связей.

1 Введение

Пользователи сети Интернет принимают активное участие в создании контента, например: оставляют комментарии в социальных сетях, пишут отзывы на товары интернет-магазинов, ведут блоги и общаются на форумах. Однако многие ресурсы дают возможность не только делиться текстовой информацией, но и оставлять персональные данные - заполнять профиль. Как правило, к таким данным относятся имя, возраст, пол, контактная информация, интересы и прочее.

Информация в профиле может оказаться неполной, например некоторые поля могут быть необязательными для заполнения, их часто оставляют пустыми. Более того, некоторые пользователи намеренно указывают неверные данные. Возникает задача автоматического предсказывания недостающих демографических атрибутов, так как, к примеру, для проведения социологических исследований, а также для работы гибридных и основанных на знаниях рекомендательных систем и таргетированной рекламы необходим наиболее полный набор характеристик [8] [1].

Во многих работах используют методы машинного обучения для решения этой задачи. Сначала осуществляют сбор данных, после чего производят обучение модели и наконец, определяют неизвестные атрибуты с помощью полученной модели и оценивают её качество [13].

В данной работе рассматривается задача предсказания демографических атрибутов пользователей социальной сети ВКонтакте; для признакового описания объектов вводятся признаки, использующие распределения подграфов в графах.

2 Постановка задачи

Распределения подграфов размера 3 и 4 в графе позволяют с высокой точностью классифицировать графы из различных доменов, таких как социальные сети, биоинформатика и графы цитирования [12]. Это даёт основания предполагать, что распределения подграфов малого размера в графе социальных связей могут быть использованы в качестве признака для описания пользователей интернет-ресурсов. Цель исследования - проверить гипотезу о том, что с помощью распределений подграфов во второй окрестности пользователя в социальном графе можно предсказывать демографические атрибуты пользователей сети ВКонтакте. Предлагается сравнить методы на основе распределений подграфов, использованных в качестве нового признака для описания пользователей, с базовым решением, основанным на рассмотрении распределений атрибутов соседей.

Для достижения поставленной цели необходимо решить следующие подзадачи:

- Изучить существующие подходы к решению поставленной задачи;
- Выбрать аккаунты пользователей ВКонтакте с полным набором известных атрибутов (пол, возраст, семейное положение, образование) для обучающей выборки;
- Для пользователей из обучающей выборки построить их вторые окрестности в графе социальных связей;
- Построить распределения подграфов в полученных окрестностях;
- Сформировать признаковые описания пользователей на основе посчитанных ранее распределений подграфов;
- Построить классификационные и регрессионные модели на основе сформированных признаковых описаний;
- Произвести сравнение качества предсказания существующих методов решения задачи.

3 Обзор существующих решений

Так как не удалось найти работы, в которых бы рассматривались основанные на использовании распределений подграфов решения задачи автоматического определения социально-демографических параметров пользователей интернет-ресурсов, в данном разделе освещены подходы к формированию признакового описания пользователей, необходимого для решения поставленной задачи с помощью методов машинного обучения, которые используют другую информацию о пользователе: текстовые сообщения, известные значения полей профиля, взаимодействие с социальной средой. Более подробно рассмотрены случаи использования графовых признаков, а также приведены описания алгоритмов на графах для решения данной задачи.

3.1 Текстовые данные

Исходными данными в задаче предсказания атрибутов являются текстовые сообщения пользователей (комментарии, отзывы, посты в личных блогах) и их профили.

Из текстов в качестве признаков можно выделить последовательности подряд идущих слов или символов. Такие последовательности называются соответственно словесными и символьными n-граммами. В работе [2] авторы используют словесные и символьные n-граммы для определения пола пользователей социальной сети Twitter. Вместе с текстами твитов - коротких сообщений в Twitter, длина которых не превышает 140 символов - рассматривают также три поля профиля: никнейм пользователя, его настоящее имя и графу "о себе". В качестве признаковой функции авторы исследования берут индикатор от n-граммы, который не равен нулю, если данная n-грамма встречается в последовательности строк, связанной с конкретным пользователем и типом текстовых данных - с твитами или одним из упомянутых выше полей профиля. Точность классификации достигает 92% при совместном рассмотрении текстов твитов и информации, указанной в полях, выбранных для постороения признакового описания.

Кроме того, из текстов можно извлекать статистические признаки, такие как средняя длина сообщений пользователя, частота знаков препинания и эмотиконов и прочие [13]. В работе [7] автор сравнивает два различных подхода к определению пола пользователей сервиса YouTube. Один из этих подходов основан на рассмотрении только текстовых данных. Из комментариев пользователей автор извлекает три группы статистических признаков: (1) символьные признаки, к которым относятся средняя длина комментария в символах, доля заглавных букв и отношение числа знаков препинания к общему количеству символов в комментариях; (2) словесные признаки, такие как средняя длина комментария в словах и отношение количества уникальных слов в комментариях к общему числу токенов; (3) признаки на основе предложений - среднее значение числа предложений в комментариях и средняя длина предложений в словах. Точность предсказания при таком подходе составляет приблизительно 90%.

3.2 Социальные взаимодействия пользователей, графовые признаки

Как уже было сказано ранее, наряду с текстовыми сообщениями пользователя анализируется его профиль. Причём рассмотрению подлежат не только значения атрибутов профиля, но и социальные связи пользователя. Социальные графы, в которых вершины соответствуют пользователям, а рёбра - наличию связей между ними, используются для формирования признакового описания.

3.2.1 Социальные представления

В работе [11] авторы решают регрессионную задачу определения возраста пользователей социальных сетей с точностью до года. Предлагаемый ими метод формирует социальные представления[1] вершин графа связей пользователей. Построенные представления используются для выделения сообществ в графе, информация о которых в дальнейшем служит ковариационным параметром.

Рассмотрим данное решение более подробно.

Для набора значений (x_i, y_i) входной и выходной переменной общее уравнение линейной регрессии имеет вид

$$y_i = w^T x_i + \epsilon_i,$$

где ϵ - случайная величина, шумовой параметр.

Когда выход y_i зависит от соседних выходных значений, модель может быть расширена с помощью дополнительных ковариационных параметров для урегулирования этих зависимостей. В случае социальных сетей граф связей между пользователями $G = (V, E)$ даёт такие зависимости. Авторы исследования [11] предлагают моделировать их с помощью социальных представлений малой размерности. С использованием метода, описанного в работе [10], строится функция отображения $\Phi : i \in V \to \mathbb{R}^{|V| \times d}$ множества вершин V на пространство социальных представлений. Задача принимает вид

$$y_i = \mathbf{w}^T \Phi(i) + \epsilon_i$$

и может быть решена методом наименьших квадратов.

[1] Социальное представление вершины графа - набор неявных признаков, которые фиксируют сходства в окрестности вершины и вхождение в определённое сообщество в графе. Эти скрытые признаки задают представление социальных отношений в непрерывном векторном пространстве с относительно небольшим числом измерений [10].

Эксперименты в исследовании [11] проводились на данных социальной сети Pokec. Средняя ошибка предсказания возраста составила 4.15 года [11].

3.2.2 Неполный расширенный социальный граф

Для решения задачи определения демографических атрибутов пользователей социальных сетей авторы работы [6] предлагают алгоритм PGPI (Partial Graph Profile Inference), который даёт возможность самостоятельно задавать число узлов социального графа для рассмотрения, что позволяет уменьшить затраты ресурсов на вычисления и достичь оптимального соотношения между количеством посещаемых вершин и точностью предсказания. Кроме того, алгоритм учитывает не только дружественные связи между пользователями, но и членство в различных сообществах, отметки "мне нравится" и просмотры, если такая информация доступна.

Условные обозначения, используемые в работе [6]:

- Социальный граф $\mathcal{G} = \{N, L, V, A\}$, где N - множество вершин этого графа (пользователи), $L \subseteq N \times N$ - множество рёбер, $V = \{V_1, V_2, \ldots, V_m\}$ - набор множеств V_i возможных значений i-ого атрибута, $A = \{A_1, A_2, \ldots, A_m\}$ - набор множеств $A_i \subseteq N \times V_i$, состоящих из пар (вершина, значение i-ого атрибута в этой вершине);

- Расширенный социальный граф $\mathcal{E} = \{N, L, V, A, G, NG, P, PG, LP, VP\}$, где N, L, V, A те же, что и в обычном социальном графе, а G - множество сообществ, $NG \subseteq N \times G$ - множество индикаторов членства пользователей в сообществах, P - множество публикаций в сообществах (изображения, тексты, видео), $LP \subseteq N \times P$ - связи между пользователями и их отметками "мне нравится"на публикациях, $VP \subseteq N \times P$ - множество связей между пользователями и просмотрами публикаций в сообществах (очевидно, $LP \subseteq VP$).

PGPI-N - вариация алгоритма PGPI, в которой используется информация о значениях атрибутов пользователей и о связях между этими пользователями.

На вход алгоритм получает социальный граф $\mathcal{G}$, вершину n_i, атрибут k, значение которого необходимо предсказать для данной вершины, и параметры $maxFacts$ и $maxDistance$ - максимальное количество используемых вершин с учётом заданной n_i и максимальная длина пути во время обхода соответственно. Выход - предсказанное значение v атрибута k для вершины n_i.

Сначала происходит инициализация множества M, состоящего из пар ключ-значение $(v; 0)$ для каждого значения v атрибута k. Затем производится обход графа $\mathcal{G}$ в ширину. Изначально в очереди Q одна вершина n_i, в множестве посещённых вершин $seen$ она же отмечена как просмотренная. Пока очередь не пуста, а количество использованных вершин меньше $maxFacts$, из Q достаётся очередная вершина n_j, подсчитывается $F_{i,j} = \frac{W_{i,j}}{dist(n_i,n_j)}$, где $W_{i,j}$ - число совпадений значений атрибутов у вершин n_i и n_j, а $dist(n_i, n_j)$ - количество рёбер в кратчайшем пути между ними. После этого $F_{i,j}$ прибавляется к значению в паре из M по ключу-значению атрибута k в вершине n_j. Если $dist(n_i, n_j) \leqslant maxDistance$, то все непосещённые смежные с n_j вершины n_h добавляются в очередь и в множестве $seen$ отмечаются как просмотренные. После окончания цикла значение атрибута k с наибольшим числом в M возвращается в качестве предсказанного для вершины n_i.

PGPI-G - вариация алгоритма PGPI, в которой используется только информация о сообществах пользователей и публикациях в них.

На вход алгоритм получает расширенный социальный граф $\mathcal{E}$, вершину n_i, предсказываемый атрибут k и параметры $maxFacts$ и $maxDistance$, где теперь $maxFacts$ - максимальное количество используемых сообществ и публикаций из G и P.

Как и в случае PGPI-N, инициализируется множество M. Пока число использованных сообществ и публикаций меньше $maxFacts$, производится обход по всем вершинам $n_j \neq n_i$ из каждого сообщества g, в котором состоят и n_j, и n_i. Для каждого такого n_j подсчитывается значение $Fg_{i,j} = W_{i,j} \times commonLikes(n_i, n_j) \times commonViews(n_i, n_j) \times \frac{commonGroups(n_i,n_j)}{|\{(n_i,x):(n_i,x)\in NG\}|} \times commonPopularAttributes(n_i, g)$,

где $commonLikes(n_i, n_j)$ и $commonViews(n_i, n_j)$ - количество понравившихся и просмотренных постов обоих пользователей n_i и n_j соответственно, $commonGroups(n_i, n_j)$ - количество общих сообществ для n_i и n_j, $commonPopularAttributes(n_i, g)$ - число атрибутов n_i, значения которых совпадают с наиболее популярными значениями атрибутов пользователей в сообществе g. Аналогично случаю PGPI-N $Fg_{i,j}$ прибавляется к значению в паре из M по ключу-значению атрибута в n_j. После того, как заканчивается цикл, ответ выбирается из M по тому же принципу, что и в PGPI-N.

Так как PGPI-N и PGPI-G структурно похожи, авторы работы [6] объединяют их в один алгоритм. При этом добавляется новый входной параметр $ratioFacts$, показывающий в каком соотношении делить максимально допустимое количество $maxFacts$ используемых "фактов" - вершин, сообществ и публикаций из N, G и P - между частями PGPI-N и PGPI-G.

PGPI-N, PGPI-G и их объединение показали более высокую точность классификации (более 90% при предсказывании пола и семейного положения), чем наивный байесовский классификатор и алгоритм распространения меток [6].

4 Исследование и построение решения задачи

4.1 Исходные данные

В качестве исходных данных был взят полученный в ИСП РАН датасет с описанием пользователей социальной сети ВКонтакте, который содержит двусторонние дружественные связи и доступные значения следующих атрибутов: пол, семейное положение, образование и возраст.

Возможные значения атрибутов:

- пол $\in$ {мужской, женский}
- семейное положение $\in$ {холост/не замужем, в браке}
- образование $\in$ {без образования, среднее, высшее}

- возраст $\in [1 \dots 110]$

Датасет - json-файл, в котором каждый пользователь представлен в виде словаря {id: <id пользователя>, friends: <список id друзей пользователя>, gender: <пол пользователя>, status: <семейное положение пользователя>, education: <образование пользователя>, age: <возраст пользователя>}.

4.2 Построение распределений подграфов

Как уже было сказано ранее, цель работы - рассмотреть возможность использования распределений подграфов в графах социальных связей в качестве признаков для описания пользователей.

По связям, указанным в датасете, для пользователя строится неориентированный социальный граф до второй окрестности $G(V, E)$, в котором $v \in V$ - это либо сам исследуемый пользователь, либо пользователь из списка его друзей или друзей друзей, а ребро $(u, v) \in E$ - дружественная связь между пользователями u и v.

Затем следует построение распределений подграфов рамера 3 с помощью библиотеки gtrieScanner[2] и размера 4 с помощью библиотеки FaSE[3] и их нормализация по норме l_2. В случае подграфов размера 4 распределения вычисляются приближённо, на каждом шаге вершина выбирается с заранее указанной вероятностью, так как точное построение слишком ресурсозатратно. Размерность векторов распределений подграфов размера 3 равна двум, так как возможных неориентированных комбинаций на трёх вершинах всего две:

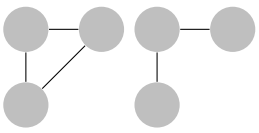

Рис. 1: Неориентированные комбинации на трёх вершинах

[2] http://www.dcc.fc.up.pt/gtries/
[3] http://www.dcc.fc.up.pt/gtries/fase/

Размерность векторов распределений подграфов размера 4 равна шести, так как возможных неориентированных комбинаций на четырёх вершинах шесть:

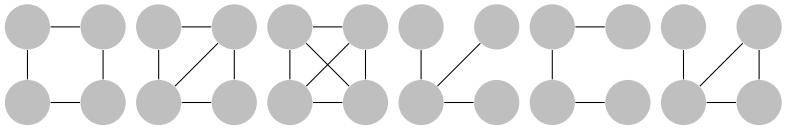

Рис. 2: Неориентированные комбинации на четырёх вершинах

4.3 Формирование признакового описания

4.3.1 Базовое решение

В качестве базового решения задачи формирования признакового описания объектов было выбрано использовать распределения признаков друзей пользователей. Для каждого из атрибутов строится свой вектор ответов Y и матрица "объектов-признаков" X:

1. Для пола $X = \begin{pmatrix} a_1 & b_1 \\ \vdots & \vdots \\ a_N & b_N \end{pmatrix}$, где a_i - количество друзей мужского пола у i-ого пользователя, b_i - женского.

2. Аналогичный вид принимает матрица X в случае семейного положения, только a_i - количество одиноких друзей у i-ого пользователя, а b_i - число друзей пользователя i, состоящих в браке.

3. Для образования $X = \begin{pmatrix} a_1 & b_1 & c_1 \\ \vdots & \vdots & \vdots \\ a_N & b_N & c_N \end{pmatrix}$, где a_i - количество друзей без образования у i-ого пользователя, b_i - со средним образованием, c_i - с высшим.

4. Для возраста $X = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{Nm} \end{pmatrix}$, где a_{ij} - количество друзей возраста j у i-ого пользователя.

После этого строки X_i нормируются по норме l_2.

4.3.2 Использование только распределений подграфов

Данный метод заключается в использовании только распределений подграфов в графах социальных связей пользователей для описания рассматриваемых объектов. Полученные способом, описанным в пункте 4.2, нормализованные по норме l_2 распределения подграфов с тремя и четырьмя вершинами формируют матрицы "объектов-признаков"

$$X = \begin{pmatrix} d_{11} & \dots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{N1} & \dots & d_{Nm} \end{pmatrix}$$

d_i - распределение подграфов в социальном графе i-ого пользователя, N - количество пользователей в выборке, m - размерность векторов, зависящая от размера подграфов (см. пункт 4.2).

4.3.3 Комбинация базового решения и распределений подграфов

В данном случае в качестве признакового описания объектов используются нормализованные по норме l_2 вектора, полученные конкатенацией распределений признаков друзей из базового решения и распределений подграфов.

$$X = \begin{pmatrix} (f_1 & | & d_1) \\ & \dots & \\ & \dots & \\ & \dots & \\ (f_N & | & d_N) \end{pmatrix}$$

f_i - распределение признаков друзей i-ого пользователя, d_i - распределение подграфов в социальном графе i-ого пользоателя, а $(f_i \mid d_i)$ - нормированный вектор-конкатенация.

4.3.4 Центроиды и распределения подграфов

Предлагаемый метод основан на использовании в качестве признакового описания конкатенаций распределений подграфов и центроидов классов, построенных с помощью распределений признаков друзей, а также позволяет перейти от многоклассовой к бинарной классификации. Подобное решение использовано для текстовой классификации [3].

Пусть дана выборка $\{(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)\}$, где ответы $y_i \in Y$. Тогда центроид класса определяется как $\overrightarrow{\mu_l} = \frac{1}{|C_l|} \sum_{i \in C_l} \overrightarrow{x_i}$, где C_l - множество индексов объектов выборки, принадлежащих классу $l \in Y$ [9].

Рассмотрим формирование матриц "объектов-признаков" X и множеств ответов Y для пола, семейного положения и образования:

- по распределениям признаков друзей для каждого значения каждого из атрибутов (то есть для каждого класса) строится его центроид;
- для каждого пользователя составляется набор векторов $(d_i \mid c_j)$, где d_i - распределение подграфов в социальном графе i-ого пользователя, $i \in \overline{1, N}$, N - количество пользователей в выборке, c_j - центроид j-ого класса рассматриваемого атрибута, $j \in \overline{1, m}$, m - количество возможных значений текущего атрибута, а также для очередного j-ого центроида в Y добавляется 1, если центроид того же класса, что и i-ый пользователь изначально, в противном случае добавляется 0;
- из полученных наборов строится X.

Задача определения возраста пользователей данным методом решается тоже как классификационная. Промежуток от 11 до 65 лет (граничные значения возраста пользователей из датасета) разбивается на 11 возрастных подгрупп по 5 лет, после чего строятся X и Y аналогично случаям пола, семейного положения и образования.

4.4 Оценка качества

Оценка качества классификации производится по F1-мере, причём рассматриваются значения F1-micro и F1-macro.

$$\text{F1-score} = 2 \times \frac{precision \times recall}{precision + recall}$$

Параметры *precision* (достоверность) и *recall* (полнота) для случая F1-micro:

$$precision = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FP_c}$$

$$recall = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FN_c}$$

где C - множество меток классов, а TP, FP и FN определяются как результаты классификации по отношению к истинным значениям.

		Истинное значение	
		positive	*negative*
Результат классификации	*positive*	TP	FP
	negative	FN	TN

Таблица 1: Определение параметров TP, FP, FN и TN

Параметры *precision* и *recall* для случая F1-macro:

$$precision = \frac{\sum_{c \in C} precision_c}{|C|}$$

$$recall = \frac{\sum_{c \in C} recall_c}{|C|}$$

где $precision_c = \frac{TP_c}{TP_c + FP_c}$, а $recall_c = \frac{TP_c}{TP_c + FN_c}$ для метки класса c.

Для регрессии используется метрика MAE (средняя абсолютная ошибка):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i^*|$$

где y_i - предсказанное занчение, y_i^* - истинное значение.

5 Описание эксперимента

5.1 Выбор инструментария

Основным языком разработки был выбран Python 3. Python - интерпретируемый, объектно-ориентированный язык программирования, преимуществами которого являются гибкость, простота и наличие большого количества специализированных библиотек для работы с данными.

Для алгоритмов классификации и регрессии была использована библиотека scikit-learn. Scikit-learn - библиотека алгоритмов машинного обучения с открытым исходным кодом [4] для языка Python. Эта библиотека была выбрана, потому что в ней реализованы все используемые в эксперименте классификаторы и алгоритмы регрессии, кроме XGBoost из одноимённой библиотеки [5], а именно: метод опорных векторов, логистическая регрессия, наивный байесовский классификатор, решающее дерево, метод k-ближайших соседей, линейная регрессия и ридж-регрессия.

Как было упомянуто ранее, для построения распределений подграфов были использованы библиотеки gtrieScanner и FaSE. Преимущество библиотеки gtrieScanner в высокой скорости работы. Библиотека FaSE была выбрана для случая подграфов размера четыре, так как в ней реализована возможность приближённого подсчёта распределений. Размер некоторых графов в выборке превышал 10^5 вершин - точное решение слишком ресурсозатратно.

5.2 Эксперимент

Структура эксперимента:

- на основе значений атрибутов, указанных в датасете, строится множество *seeds*, которое содержит пользователей с полностью заполненным профилем;
- для каждого пользователя из множества *seeds* строится вторая окрестность в социальном графе, после чего следует построение распределений подграфов размера три и четыре (см. пункт 4.2);
- с помощью полученных распределений подграфов формируются матрицы "объектов-признаков" методами, описанными в пункте 4.3;
- методом скользящего контроля (в данном эксперименте разбиение на 5 частей) производится оценка эффективности классификационных и регрессионных моделей (метода опорных векторов, логистической регрессии, наивного байесовского классификатора, решающего дерева, метода k-ближайших соседей, линейной и ридж-регрессии) на полученных вариантах признакового описания объектов.

Всего		7967
Пол	мужской	4152
	женский	3815
Семейное положение	в браке	4768
	холост/не замужем	3199
Образование	без образования	922
	среднее	678
	высшее	6467

Таблица 2: Распределение значений атрибутов пользователей из множества *seeds*

5.3 Результаты эксперимента

В задачах классификации были использованы реализации следующих алгоритмов для предсказывания значений атрибутов: LinearSVC, LogisticRegression, GaussianNB, DecisionTree, KNN из библиотеки scikit-learn и XGBoost из одноимённой библиотеки.

В случае предсказывания **пола** наилучший результат на всех предложенных признаковых описаниях был получен классификатором XGBoost.

	XGBoost	
	F1-micro	F1-macro
Распределения признаков друзей	**0.731089**	**0.730340**
Распределения подграфов размера 3	0.533174	0.508858
Распределения подграфов размера 4	0.545602	0.531960
Признаки друзей и подграфы размера 3	**0.740190**	**0.732807**
Признаки друзей и подграфы размера 4	**0.716905**	**0.712217**
Подграфы размера 3 и центроиды	0.533174	0.549220
Подграфы размера 4 и центроиды	0.536928	0.536927

Таблица 3: Значения F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении пола

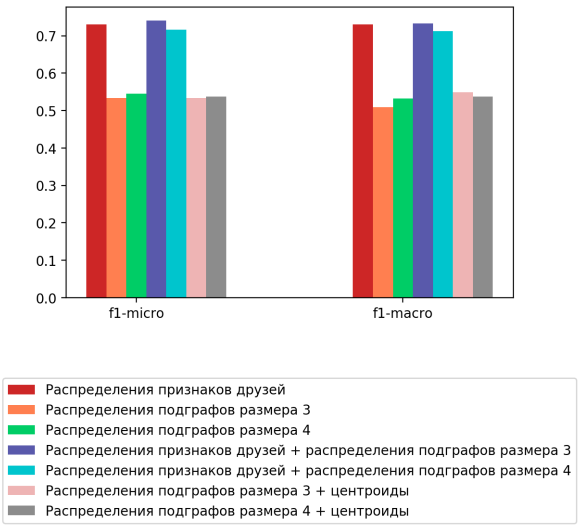

Рис. 3: F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении пола

Значения F1-micro и F1-macro превышают 70% при использовании распределения атрибутов друзей в качестве признака для описания пользователей. Показатели, полученные на распределениях подграфов, эквивалентны результатам случайного угадывания: классификатор выбирает наиболее многочисленный класс и тем самым обеспечивает наибольшую вероятность попадания в правильный ответ.

В случае определения **семейного положения** пользователей наилучшие показатели также у классификатора XGBoost.

	XGBoost	
	F1-micro	F1-macro
Распределения признаков друзей	**0.701546**	**0.695792**
Распределения подграфов размера 3	0.613874	0.587709
Распределения подграфов размера 4	0.602699	0.585391
Признаки друзей и подграфы размера 3	**0.696813**	**0.686818**
Признаки друзей и подграфы размера 4	**0.677957**	**0.666956**
Подграфы размера 3 и центроиды	0.610526	0.610504
Подграфы размера 4 и центроиды	0.599056	0.598250

Таблица 4: Значения F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении семейного положения

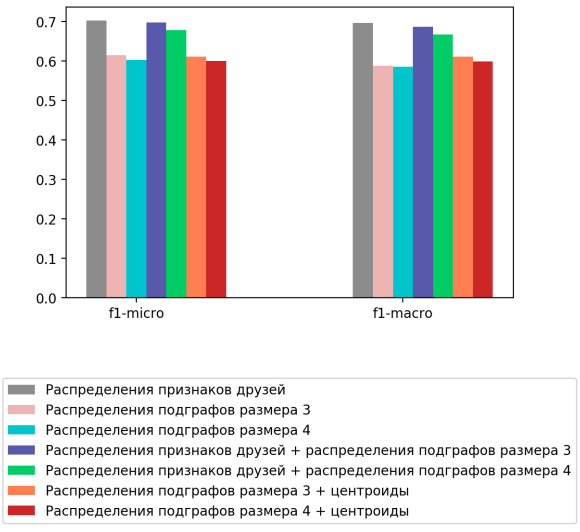

Рис. 4: F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении семейного положения

Как и в случае с предсказыванием пола, наиболее высокие значения F1-micro и F1-macro принимают при испольовании распределения признаков друзей для описания пользователей. Результаты использования в качестве признака распределений подграфов эквивалентны результатам случайного угадывания.

Для **образования** наилучшие результаты были получены методом опорных векторов на всех рассмотренных вариантах формирования описания объектов.

	LinearSVC	
	F1-micro	F1-macro
Распределения признаков друзей	0.812910	0.466753
Распределения подграфов размера 3	0.791388	0.342256
Распределения подграфов размера 4	0.761307	0.362939
Признаки друзей и подграфы размера 3	0.818178	0.686818
Признаки друзей и подграфы размера 4	0.765739	0.627155
Подграфы размера 3 и центроиды	0.860925	0.843581
Подграфы размера 4 и центроиды	0.834517	0.815764

Таблица 5: Значения F1-micro и F1-macro на классификаторе LinearSVC для различных признаковых описаний при определении образования

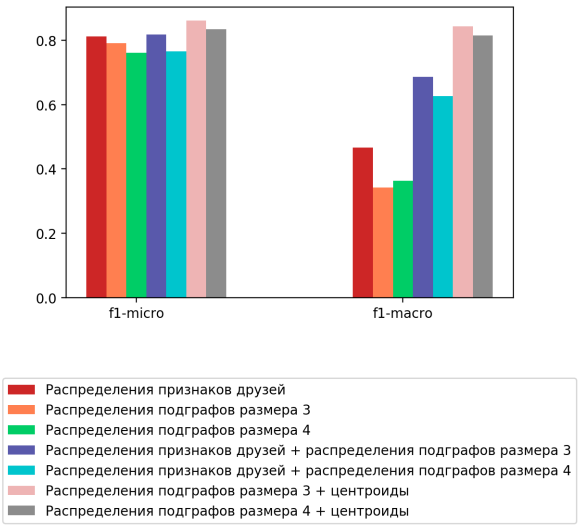

Рис. 5: F1-micro и F1-macro на классификаторе LinearSVC для различных признаковых описаний при определении образования

Выборка, составляющая множество *seeds*, неоднородна по образованию, чем объясняется разница между значениями F1-micro и F1-macro. Самые низкие показатели в случае использования только распределения подграфов для описания пользователей, однако даже в этом случае F1-micro превышает 75%. Причиной этого является то, что примерно такую долю множества *seeds* составляют пользователи одного класса (пользователи с высшим образованием).

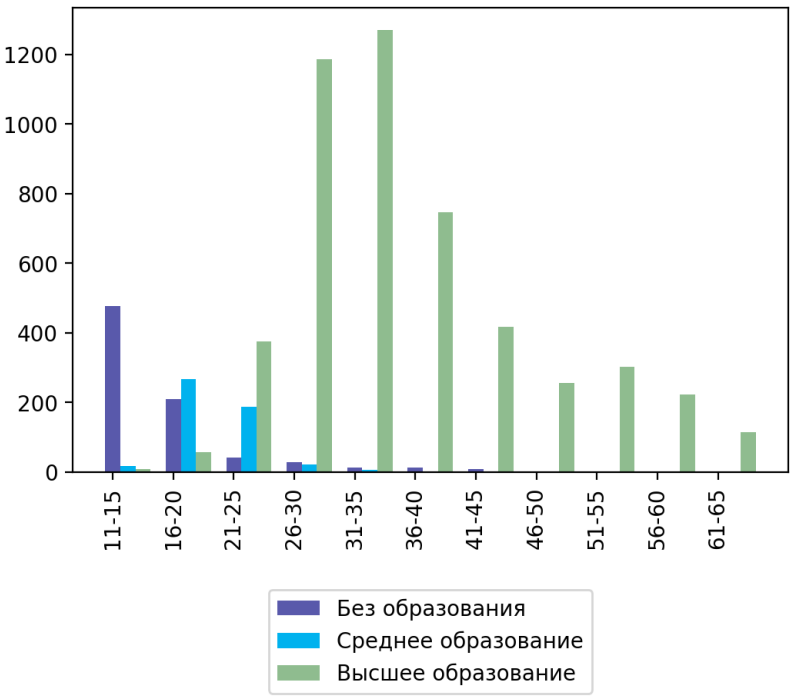

Рис. 6: Распределение уровня образования по возрастным группам

Из распределения, представленного на рис. 4 можно сделать предположение, что полностью заполняют профиль в основном люди среднего возраста (26 - 40 лет) с высшим образованием.

При решении задачи определения **возраста** пользователей с признаковым описанием, которое использует распределения подграфов и центроиды классов, наилучший результат показал классификатор XGBoost. Этот метод формирования признакового описания даёт значительное улучшение качества предсказания по сравнению с базовым решением, однако вариация такого подхода с заменой векторов распределений подграфов на вектора распределений

признаков друзей приводит к ещё большему росту показателей. Исполь-зование центроидов классов даёт настолько весомый прирост значений F1-micro и F1-macro именно в задаче предсказывания возраста пользователей, потому что в этой задаче значителен переход от многоклассовости к бинарной классификации (изначально рассматривается 11 классов - интервал от 11 до 65 лет, разбитый на возрастные подгруппы по 5 лет), который как раз осуществляется при использовании центроидов изначальных классов (см. пункт 4.3.4).

	XGBoost	
	F1-micro	F1-macro
Распределения признаков друзей	0.519502	0.466896
Распределения подграфов размера 3 и центроиды	**0.909091**	**0.539099**
Распределения признаков друзей и центроиды	**0.913499**	**0.560735**

Таблица 6: Значения F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении возраста

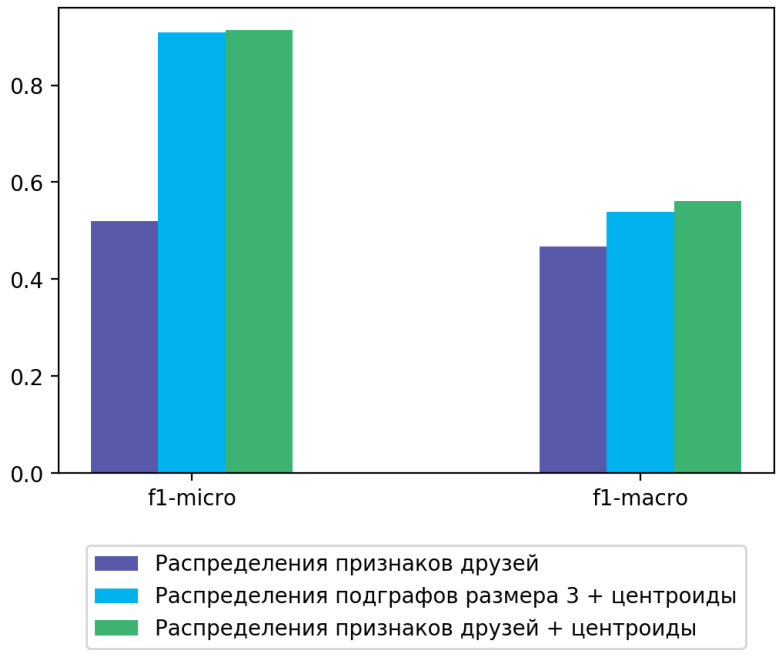

Рис. 7: F1-micro и F1-macro на классификаторе XGBoost для различных признаковых описаний при определении возраста

На регрессионной задаче предсказывания **возраста** наилучший результат

показал алгоритм ридж-регерссии для всех рассматриваемых вариантов признакового описания.

	RidgeRegression
	MAE
Распределения признаков друзей	**4.539769**
Распределения подграфов размера 3	7.081688
Распределения подграфов размера 4	7.704280
Признаки друзей и подграфы размера 3	**4.549047**
Признаки друзей и подграфы размера 4	**4.837654**

Таблица 7: Средняя абсолютная ошибка определения возраста алгоритмом RidgeRegression

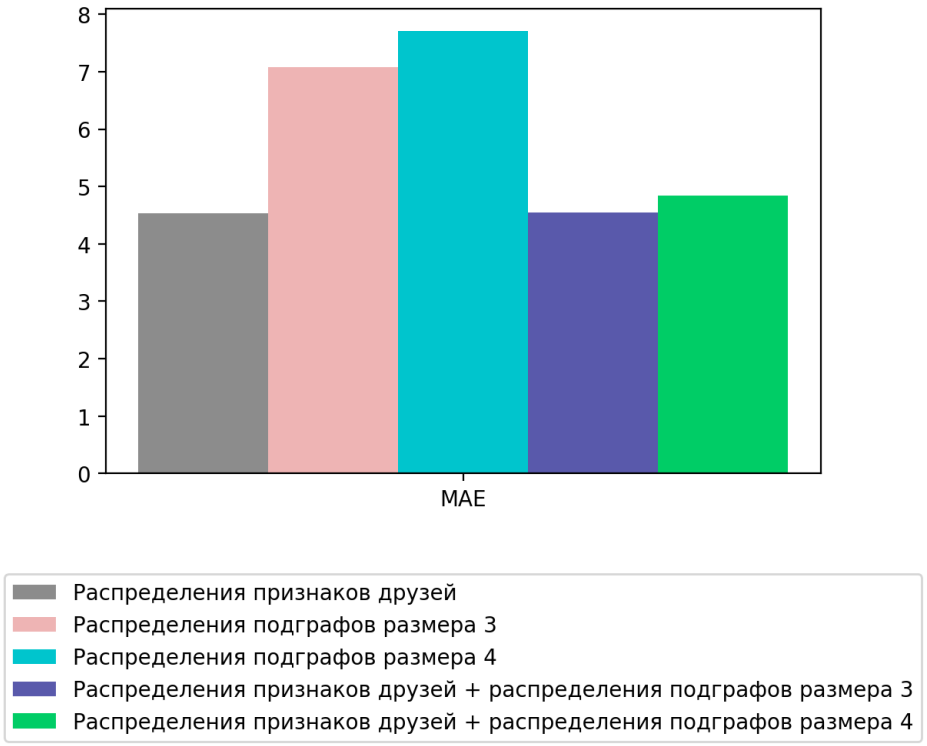

Рис. 8: Средняя абсолютная ошибка определения возраста алгоритмом RidgeRegression

Средняя абсолютная ошибка оказалась минимальна в случае базового решения задачи формирования описания объектов.

6 Заключение

В рамках квалификационной работы:

- были изучены существующие подходы к решению задачи автоматического определения демографических атрибутов пользователей социальных сетей;
- для решения данной задачи было разработано и реализовано несколько методов с использованием распределений подграфов в графах социальных связей в качестве нового признака для описания объектов;
- была проведена экспериментальная оценка эффективности методов на датасете с описанием пользователей социальной сети ВКонтакте.

В ходе исследования обнаружить зависимость между социально-демографическими параметрами пользователей и распределениями видов подграфов в их неориентированных графах социальных связей не удалось. В дальнейшем можно рассмотреть не только графы социальных связей, но и графы социальных взаимодействий (членство в сообществах, просматриваемые посты и прочее), что, возможно, приведёт к более точному предсказанию значений атрибутов пользователей.

Список литературы

[1] Sahami M. Bharat K. Lawrence S. «Generating user information for use in targeted advertising». В: (2003).

[2] John D. Burger и др. «Discriminating Gender on Twitter». В: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, с. 1301—1309. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145568.

[3] *Code for Large Scale Hierarchical Text Classification competition*. URL: https://github.com/nagadomi/kaggle-lshtc.

[4] *Code for scikit-learn library*. URL: https://github.com/scikit-learn/scikit-learn.

[5] *Code for XGBoost library*. URL: https://github.com/dmlc/xgboost.

[6] Raïssa Yapan Dougnon, Philippe Fournier-Viger и Roger Nkambou. «Inferring User Profiles in Online Social Networks Using a Partial Social Graph». В: *Advances in Artificial Intelligence*. Под ред. Denilson Barbosa и Evangelos Milios. Cham: Springer International Publishing, 2015, с. 84—99. ISBN: 978-3-319-18356-5.

[7] Katja Filippova. «User Demographics and Language in an Implicit Social Network». В: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, 2012, с. 1478—1488. URL: http://dl.acm.org/citation.cfm?id=2390948.2391117.

[8] Qing Li и Byeong Man Kim. «Constructing User Profiles for Collaborative Recommender System». В: *Advanced Web Technologies and Applications*. Под ред. Jeffrey Xu Yu и др. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, с. 100—110. ISBN: 978-3-540-24655-8.

[9] *Nearest centroid classifier*. URL: https://en.wikipedia.org/wiki/Nearest_centroid_classifier.

[10] Bryan Perozzi, Rami Al-Rfou и Steven Skiena. «DeepWalk: Online Learning of Social Representations». В: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: ACM, 2014, с. 701—710. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623732. URL: http://doi.acm.org/10.1145/2623330.2623732.

[11] Bryan Perozzi и Steven Skiena. «Exact Age Prediction in Social Networks». В: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: ACM, 2015, с. 91—92. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742765. URL: http://doi.acm.org/10.1145/2740908.2742765.

[12] A. Wegner. *Random graphs with motifs*. Preprint. Max Planck Institute for Mathematics in the Sciences, 2011. URL: http://www.mis.mpg.de/de/publications/preprints/2011/prepr2011-61.html.

[13] Кузнецов С.Д. Гомзин А.Г. «Методы построения социо-демографических профилей пользователей сети Интернет». В: *Труды ИСП РАН* (2015). DOI: 10.15514/ISPRAS-2015-27(4)-7.