МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РЕАЛИЗАЦИЯ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ СОЦИАЛЬНЫХ СЕТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Балашовой Татьяны Алексеевны

Научный руководитель:
зав.кафедрой, к.ф.-м.н., доцент _____ Огнева М.В.

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М.В.

подпись, дата

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы.

Задачи машинного обучения находят очень широкое применение в различных сферах человеческой жизни: медицине, спорте, музыке, социальной среде и т.д.

Одним из видов машинного обучения является обучение без учителя. Задачи такого обучения актуальны, поскольку далеко не всегда с уверенностью можно сказать, что данные, которые необходимо анализировать, окажутся маркированными. Одной из важных задач обучения без учителя является задача кластеризации, которая заключается в разбиении множества объектов на сообщества.

На данный момент существует множество алгоритмов решения задачи кластеризации, но не всегда очевидно, какой лучше использовать в конкретной ситуации. Не существует универсального алгоритма: каждый имеет свои преимущества, недостатки и ограничения.

Графы являются очень важным объектом в научных исследованиях. С их помощью можно представить огромное количество информации. Социальные сети, дорожные карты, взаимодействия белков, ссылки в интернете – это только лишь некоторые примеры их применения.

Задача поиска сообществ в неориентированных графах является одной из задач кластеризации, которая связывает машинное обучение с теорией графов. Данная задача возникает, например, тогда, когда нужно выделить дружественные сообщества по интересам, разбить страны мира на группы схожих по экономическому положению государств или по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества.

Цель бакалаврской работы – реализация и проведение сравнительного анализа различных алгоритмов кластеризации социальных сетей на примере реальных прикладных задач.

Поставленная цель определила **следующие задачи**:

1. Дать определения понятиям, относящимся к машинному обучению и теории графов.
2. Разобрать основные методы решения задачи кластеризации социальных сетей и оценки их качества.
3. Подобрать и подготовить данные для внедрения в реализацию.
4. Реализовать разобранные методы и оценить их качество самостоятельно и с помощью методов библиотеки.
5. Провести сравнительный анализ методов на примерах подготовленных данных (тестовых и реальных).

Методологические основы проблемы поиска сообществ представлены в работах Воронцова, Судакова, Муромцева.

Практическая значимость бакалаврской работы заключается в реализации метрик качества кластеризации и в работе с не обезличенными, реальными данными социальной сети ВКонтакте.

Структура и объём работы. Бакалаврская работа состоит из введения, 9 разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы – 55 страниц, из них 40 страниц – основное содержание, включая 12 таблиц, список использованных источников информации – 21 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Машинное обучение: основные понятия» посвящен основам машинного обучения и областям его применения.

Обучение с учителем – один из разделов машинного обучения, посвященный решению следующей задачи.

Имеется множество объектов (ситуаций), множество возможных ответов (откликов, реакций) и некоторая неизвестная зависимость между объектами и ответами. Известна только конечная совокупность прецедентов – пар «объект, ответ», называемая обучающей выборкой. На основе этих данных требуется восстановить зависимость, то есть построить алгоритм, способный для любого объекта выдать достаточно точный ответ.

К задачам обучения с учителем относят задачи классификации, регрессии и ранжирования.

Обучение без учителя – раздел машинного обучения, изучающий широкий класс задач обработки данных, в которых известны только описания множества объектов из обучающей выборки, и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

К задачам обучения без учителя относятся задачи кластеризации, обобщения, поиска правил ассоциации, сокращения размерности, визуализации данных.

Второй раздел «Кластеризация: общая постановка задачи» посвящен введению формальной поставки задачи кластеризации в целом. Она выглядит следующим образом.

Пусть X – множество объектов, Y – множество номеров кластеров. Задана функция расстояния между объектами $\rho\,(x,\ x')$. Имеется конечная обучающая выборка объектов $X_m = \{x_1, x_2, .., x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

Третий раздел «Кластеризация графов: постановка задачи» посвящен основам теории графов и формальной постановке задачи кластеризации графов.

Под графом понимается упорядоченная пара $G = (V, E)$, где V – непустое множество объектов, называемых вершинами, $E \subset V^2$ – множество ребер.

Графы могут быть ориентированными и неориентированными.

Ориентированный граф представляет собой упорядоченную пару

$$G = (V, E): \forall (u, v) \in E \ \exists (v, u) \in E$$

Неориентированный граф представляет собой упорядоченную пару

Формальная постановка задачи кластеризации графов.

Дан неориентированный граф $G = (V, E)$, где V – множество вершин, E – множество ребер. Необходимо получить покрытие множества вершин, называемое кластеризацией, которое выглядит следующим образом:

$V = \bigcup_{C^i \in C} C^i$, где $C^i \neq \emptyset$ – покрытие множества вершин кластера с номером i, C – покрытие множества всех вершин.

Четвертый раздел «Метрики оценки качества» посвящен описанию различных метрик оценки качества кластеризации – внутренних и внешних.

Внешние методы – методы, основанные на сравнении результата с априори известным разделением на классы, так называемыми, ground-truth сообществами.

Внутренние методы – методы, определяющие качество кластеризации только по признаковым описаниям объектов, без итогового разбиения.

Несмотря на то, что на данный момент существует множество различных методов оценки, проблема оценки качества кластеризации трудноразрешима в силу того, что не существует оптимального алгоритма кластеризации и многие из них не способны определить истинное количество кластеров в данных.

Пятый раздел «Теоретические описания алгоритмов» посвящен особенностям реализаций рассматриваемых алгоритмов, а также их вычислительной сложности.

Алгоритм Infomap использует механизм случайных блужданий: задача поиска сообществ в графе сводится к минимизации длины кода пути, который проделает «случайный блуждатель».

Алгоритм Labelpropagation основывается на том принципе, что вершина относится к тому сообществу, что и большинство ее соседей.

Целью алгоритма Fastgreedy является жадная оптимизация модулярности.

Алгоритм Walktrap, также как и Infomap, обращается к случайным блужданиям. Основная идея алгоритма заключается в том, что короткие случайные блуждания не приводят к выходу из сообщества.

Лувенский алгоритм кластеризации графов, также, как и Fastgreedy, основан на жадной оптимизации модулярности, однако отличается тем, что зависит от перебора вершин на этапе оптимизации модулярности.

Шестой раздел «Анализ данных социальной сети ВКонтакте» посвящен описанию специального модуля для создания скриптов для социальной сети ВКонтакте – vk_api. С его помощью можно полноценно пользоваться функционалом данной социальной сети – отправлять сообщения, просматривать фотографии, списки друзей, аудиозаписей, публиковать записи на стене и т.д.

Для совершения этих манипуляций в данном модуле существуют определенные методы, которые подразделяются в зависимости от объектов сети, с которыми они работают, на следующие категории: аккаунт, друзья, реклама, базы данных, документы, аудиозаписи и т.д.

В данной работе модуль vk_api рассматривается как инструмент для осуществления сбора данных с последующей их обработкой и анализом (парсинга).

Седьмой раздел «Расчет оценок качества кластеризации» посвящен

реализации алгоритмов расчета метрик качества. Были рассчитаны внешние метрики качества (индекс Rand, индекс Фоулкса-Мэллова), а также внутренние (индекс WSS, индекс BSS, силуэт).

Восьмой раздел «Построение дружественного графа социальной сети ВКонтакте» посвящен реализации метода, инициализирующего граф социальной сети ВКонтакте.

Модель графа сети можно описать следующим образом:

1) Вершины графа – пользователи социальной сети ВКонтакте, которые находятся в списке друзей конкретного пользователя.
2) Если Пользователь А и Пользователь Б являются друзьями, то между вершиной, соответствующей идентификатору Пользователя А, и вершиной, соответствующей идентификатору Пользователя Б, проводится неориентированное ребро.

Особенность реализации данного метода заключается в том, что при получении списка страниц друзей пользователя какие-то из них могут оказаться закрытыми, замороженными, заблокированными, удаленными. Как следствие, нельзя получить доступ к информации данной страницы. В таком случае вершина в графе, соответствующая этой странице, будет изолирована.

Девятый раздел «Сравнение результатов работы алгоритмов поиска сообществ» посвящен проведению сравнительного анализа различных алгоритмов кластеризации сетей.

В первой подчасти данного раздела анализ проводился на искусственных, обезличенных данных, чье внутреннее содержание, как следствие, нельзя посмотреть.

Датасеты для этих примеров были взяты из коллекции наборов данных больших сетей Стэнфордского университета. Результаты для социальной сети Facebook без информации об итоговом разбиении показали, что количество кластеров у всех моделей алгоритмов разное, что может вызвать затруднения в реальной оценке качества. Это вполне ожидаемый результат

поскольку алгоритмы могут делить по-разному. Во втором примере с уже имеющейся информации о разбиении для одной трети объектов не было указано ground-truth сообщество. Результаты для этой сети говорят о том, что что между собой разбиения в целом схожи, однако с ground-truth сообществами не очень хорошо совпадают. Это связано с тем, что изначально в данных были пропуски – не для каждого объекта было известно его истинное сообщество. Если решать данную проблему, удаляя из графа те вершины, для которых неизвестно ground-truth сообщество, то меняется вся структура графа, а, следовательно, рассматриваемый пример будет уже не продуктовой сетью, а чем-то другим.

Во второй подчасти данного раздела анализ проводился на реальных данных социальной сети ВКонтакте. Был взят пользователь, а список всех его друзей был разбит на ground-truth сообщества – одногруппники, одноклассники, спортсмены и т.д.

Опираясь на полученные результаты, можно сказать то, что в целом разбиения не похожи на ground-truth сообщества. Если бы данные были бы «обезличены», можно было бы на этом остановиться и сделать вывод о том, что это просто неудачный пример – либо пользователь не так подобран, либо отношение нужно было задать по-другому. Однако, поскольку есть доступ к их внутреннему содержанию, имеет смысл проанализировать частные случаи сообществ с точки зрения самой информации в них.

При анализе получившихся сообществ можно использовать в качестве основной информации исходное разбиение на ground-truth сообщества, а также разбиения, являющиеся результатами работы алгоритмов. Объеденив эти два подхода, можно сделать вывод о том, что по времени выполнения самыми быстрыми являются Lovain и Labelpropagation – на небольших данных работают примерно одинаково, однако с увеличением объема данных Lovain начинает выигрывать. Если же говорить о разбиениях, то несмотря на то, что каждый делит по-своему, получаются вполне осмысленные группы.

То, какой вариант разбиения лучше, зависит от условий решаемой в данный момент задачи.

ЗАКЛЮЧЕНИЕ

В данном исследовании были рассмотрены пять алгоритмов решения задачи поиска сообществ в неориентированных графах (Infomap, Labelpropagation, Fastgreedy, Walktrap, Lovain), проведен их сравнительный анализ на тестовых и реальных данных.

Был разработан алгоритм построения графа социальной сети Вконтакте, который затем вручную был разбит на ground-truth сообщества в соответствии с определенными критериями.

Была изучена внутренняя структура основных ground-truth сообществ, используя разбиения алгоритмов, а также сообществ, определяемых алгоритмами, на составленные ground-truth сообщества.

Следует отметить, что очень важно иметь доступ к внутреннему содержанию предоставляемых данных, поскольку без этого нельзя объективно оценить качество кластеризации, важно получать не только общие результаты, но и анализировать частные случаи.

Обобщив все полученные результаты, можно сделать вывод о том, что задача поиска сообществ в неориентированных графах до сих пор является открытой и не до конца изученной. Для ее решения на данный момент существует много алгоритмов, и в конкретном случае каждый из них ведет себя по-своему, имеет свои преимущества, недостатки и ограничения.

Основные источники информации:

1. Машинное обучение [Электронный ресурс] – URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5

2. Обучение с учителем [Электронный ресурс] – URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC

3. Обучение без учителя [Электронный ресурс] - URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B1%D0%B5%D0%B7_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D1%8F

4. Судаков С.А. Кластерный анализ в психиатрии и клинической психологии. – Казань: Издательство «Медицинское информационное агентство», 2010. – С.23-25.

5. Муромцев. Методы анализа социальных графов и поиска сообществ [Электронный ресурс] – URL: https://www.slideshare.net/msucsai/ss-57225859

6. Stanford Large Network Dataset Collection [Электронный ресурс] – URL: https://snap.stanford.edu/data/