К вопросу выбора методов предиктивного анализа данных социальных медиа

И.С. Калытюк, Г.А. Французова, А.В. Гунько

ФГБОУ ВО Новосибирский государственный технический университет, Новосибирск, Россия

Аннотация. В данной статье обсуждается вопрос выбора методов предиктивного анализа в задачах, связанных с анализом данных, полученных из социальных медиа. По мере развития сети интернет, а также социальных медиа, более простым стал доступ и распространение информации, ведь сами пользователи сети являются одновременно создателями и получателями различной информации. Основная разновидность социальных медиа - социальные сети. Из наиболее известных можно выделить следующие: *Facebook, VK, Instagram, YouTube, Twitter,* Одноклассники, а также мессенджеры *WhatsApp, Telegram*. Важнейшие функции социальных медиа – влияние на восприятие, отношение и конечное поведение потребителей. В основе предиктивной аналитики лежит автоматический поиск связей, аномалий и закономерностей между различными факторами. Данный метод анализа данных социальных медиа находится на стадии своего развития. Отсюда очевидна актуальность исследования методов предиктивной аналитики на примере анализа данных социальных медиа, извлеченных из открытых источников. Практическую ценность результаты предиктивной аналитики могут представлять в следующих случаях: создание рекомендательных сервисов для заинтересованных пользователей в определенном товаре/услуге, прогнозирование аварий/сбоев, принятие оптимальных решений. Научной новизной является оценка применимости методов предиктивной аналитики в области анализа данных социальных медиа. Рассматриваются возможные методы контролируемого и неконтролируемого обучения, такие как регрессия, классификация, кластеризация. Описываются подробно такие методы регрессии, как множественная линейная регрессия, полиномиальная регрессия, регрессионные деревья. Из методов классификации выделены следующие известные методы: логистическая регрессия, деревья классификации, поддерживающие векторные машины, случайные леса, *k*-ближайшие соседи, наивный байесовский классификатор. В области кластеризации внимание уделяется методам иерархической кластеризации, кластеризации методом *k*-средних, анализа основных компонентов, методам кластеризации и ранжирования, использующихся в теории графов. Выделены алгоритмы, наиболее адекватные для данной предметной области, которые могут привести к получению новой информации, полезной для разработчиков и пользователей.

Ключевые слова: предиктивный анализ, социальные медиа, регрессия, классификация, кластеризация.

ВВЕДЕНИЕ

Предиктивная аналитика (также называемая в литературе прогнозной, либо предсказательной) – комплекс методов и алгоритмов, использующихся в анализе данных. Основная задача, решаемая в данной области – прогнозирование последующего поведения определенных объектов и субъектов с целью решения задач принятия наилучшего решения. В процессе обработки и анализа данных выделяются факторы, которые приводят к получению новой (спрогнозированной) информации [1].

Данные методы широко используются в бизнесе при решении следующих задач:

- предложение дополнительных товаров и услуг на основе уже приобретенных (или планируемых к приобретению);
- оптимизация определенных параметров оборудования на производстве (например, изменение температуры оборудования на основе предполагаемых погодных условий, степени износа и времени непрерывной работы);
- работа с клиентами, учитывающая статистику – выдача кредитов в банковской сфере на основании сегментации по определенным признакам;
- управление рисками на предприятии – учет тех статистических данных, при которых сотрудники увольняются и создание условий для того, чтобы не допустить нехватку рабочей силы [2].

Для предиктивного анализа обычно используются большие массивы данных. Данная информация часто является неструктурированной. Для сортировки полученных данных предназначены такие методы, как классификация и кластеризация.

На этапе создания предиктивной модели необходимо:

- поставить решаемую задачу – какие знания предпочтительно получить на выходе, временные интервалы актуальности этих знаний;
- выбрать модель для решения (статистическую/математическую) – по каким факторам можно провести адекватный анализ, учитывая вес каждого из них [3].

Целью данной статьи является выбор наиболее адекватных методов, которые могут использоваться для предиктивного анализа данных социальных медиа.

ОСНОВНЫЕ МЕТОДЫ ПРЕДИКТИВНОГО АНАЛИЗА

Так как предиктивный анализ связан с машинным обучением, можно выделить основные типы аналитики:

- контролируемое обучение;
- неконтролируемое обучение.

Контролируемое обучение, в свою очередь, можно разделить на следующие 2 категории: регрессия для количественных ответов и классификация для бинарных ответов.

Регрессия получила достаточно широкое применение. В случае данных методов выбирается количественная переменная – предпочтительный результат анализа. К примеру, цена покупки компьютера основывается на следующих предикторах: объем оперативной памяти, объем видеопамяти, частота процессора и другие. Связь между предполагаемой ценой и предикторными переменными – основа предиктивной модели. Также необходимо задать каждому из предикторов свой вес – коэффициент влияния на результат.

В машинном обучении и анализе данных основными являются такие типы регрессии, как линейная, полиномиальная регрессии, регрессионные деревья. Эти методы просты в использовании и эффективны, однако, в некоторых случаях могут использоваться и другие, в зависимости от задачи [4].

Второй большой категорией контролируемого обучения является классификация. Классификация позволяет разбивать данные на несколько категорий, например, низкая, средняя или высокая температура на улице. В этом случае, анализ происходит не только по предиктовым переменным, но и по уже известным выходным данным. Исследуются различные комбинации предикторов, находится наиболее влияющие на выходной результат – данный набор данных является обучающей выборкой. Данные наблюдения далее переносятся на тестовый набор данных, по которым уже и происходит непосредственно классификация.

Известны следующие методы классификации: логистическая регрессия, деревья классификации, поддерживающие векторные машины, случайные леса, *k*-ближайшие соседи, наивный байесовский классификатор.

Множественная линейная регрессия

Один из самых известных методов линейной регрессии – множественная линейная регрессия. В данном методе есть определенное количество входных независимых параметров и один выходной зависимый [5].

Выходной зависимый параметр является линейной комбинацией входных независимых параметров. Коэффициенты при независимых параметрах и смещение обычно вычисляются при помощи алгоритма стохастического градиентного спуска [6]. При его использовании в машинном обучении коэффициенты изменяются после каждого из объектов обучения. В сравнении со стандартным градиентным спуском, скорость работы стохастического спуска выше при обработке больших массивов данных.

Несмотря на простоту моделирования и широкую применимость в задачах с несложными зависимостями и малым количеством данных, важно помнить, что при линейной регрессии может возникнуть большая чувствительность к выбросам, что не всегда приводит к корректным результатам [7].

Полиномиальная регрессия

Также возможно использование такого метода, как полиномиальная регрессия. Она используется в случаях нелинейно разделяемых данных [8].

Главное отличие от линейной множественной регрессии – значение степени весовых коэффициентов может быть более 1 [9].

Для выбора степени нужны определенные знания о взаимосвязи входных и выходных параметров. При неправильном выборе степеней возможно такое явление, как перенасыщение модели.

Регрессионные деревья

Помимо этих методов существуют еще регрессионные деревья. Это один из двух подтипов деревьев решений, при котором возвращается результат, являющийся вещественным числом [10].

Дерево состоит из «листьев» (значений целевой функции) и «веток» (атрибуты зависимости). Спуск по дереву происходит по определенному алгоритму. Существует несколько алгоритмов, которые широко применяются в анализе данных, такие как *CART, ID*3 и его модификация *C*4.5, *MARS* [11].

Все эти алгоритмы работают в одном ключе: данные разделяются на две группы, схожие элементы оказываются в одной из групп. Процесс спуска продолжается для каждой группы. По мере разделения, в каждом листе остается все меньше и меньше данных, которые постепенно становятся более однородными. Данный метод разбиения носит название рекурсивного деления, и, в общем случае, состоит из 2 шагов:

1. Постановка бинарного вопроса, который наиболее подходит для последующего разделения элементов на однородные группы;
2. Повторение шага №1 до момента достижения критерия остановки.

Критерии остановки могут быть различными:

- остановка при превышении «лимита роста» (в случае задания его при постановке задачи);
- остановка, если все данные в листе присвоены одному из классов;
- дальнейшие бинарные вопросы не улучшают однородность данных.

Так как регрессионные деревья – это деревья решений, они имеют следующие недостатки: нестабильность при изменении изначальных данных (добавление новых данных влияет на вид всего дерева – восприимчивость к переобучению), неточность (выбор лучшего бинарного вопроса не всегда приводит к точному прогнозированию).

Логистическая регрессия

Одним из широко применяемых методов классификации является логистическая регрессия. В данном методе выходной параметр – вероятность события, учитывающая множество входных параметров [12]. В случае логистической

регрессии возникает следующая проблема: выходной параметр должен находится в промежутке от 0 до 1, о чем модель не знает. То есть, необходимо ввести ограничения на выходную переменную [13]. Для этого используется уравнение регрессии (логит-преобразование).

При расчете коэффициентов применяются различные градиентные методы, например, метод сопряженных градиентов, методы переменной метрики.

Основная идея логистической регрессии схожа с задачей линейной регрессии – разделение исходного пространства линейной границей на две части. Данная граница называется линейным дискриминантом.

Деревья классификации

Вторым из подтипов деревьев решений являются деревья классификации. Они аналогичны регрессионным деревьям, за исключением того момента, что деревья классификации на выходе возвращают результат, являющийся символьным (принадлежность к определенному классу).

Классификационные деревья имеют те же недостатки: нестабильность при изменении изначальных данных (добавление новых данных влияет на вид всего дерева – восприимчивость к переобучению), неточность (выбор лучшего бинарного вопроса не всегда приводит к точному прогнозированию).

Для того, чтобы обойти эти ограничения при задачах классификации, можно пренебречь тем, что следует использовать лучшее разбиение данных. Возможно использование нескольких вариантов деревьев для получения более точных и постоянных результатов – комбинирование прогнозов. Метод, использующийся при данной постановке задачи, носит название случайного леса.

Поддерживающая векторная машина

В машинном обучении также известен такой метод, как поддерживающая векторная машина (машина опорных векторов – *SVM*). В литературе также можно встретить следующее название: классификатор с максимальным зазором.

Данный метод позволяет реализовать бинарную классификацию, то есть разделение входных векторов на две составляющие (да-нет, положительный-отрицательный и другие). Также, как и в случае всех алгоритмов классификации, присутствует обучающая и тестовая выборка. При обучении находятся значения, лежащие на границе подмножеств, и уже по этим значениям строится граница классификации. Эти значения называются опорными векторами.

Поддерживающая векторная машина находит гиперплоскость разделения, которая обеспечивает максимальный зазор между значениями, находящихся в разных классах [14]. Для этого изначальный вектор переносится в пространство большей размерности – при помощи функции ядра.

В своей основе поддерживающая векторная машина обеспечивает линейную классификацию. Для решения задачи нелинейной классификации требуется использовать нелинейные ядра [15], например, однородный полином, неоднородный полином, гауссовскую функцию радиального базиса.

Случайный лес

Достаточно часто используется такой метод классификации, как случайный лес. Идея заключается в построении некоторого количества деревьев решений, называемого ансамблем. Полное определение ансамбля – прогностическая модель, полученная при помощи комбинирования предсказания от других моделей. Модели в этом случае должны быть независимы друг от друга, взаимно не коррелировать. Это позволяет обойти те ограничения, которым подвержены деревья решений (регрессионные деревья и деревья классификации) – нестабильность и неточность. Бинарные вопросы в случае случайного леса выбираются для каждого из деревьев случайным образом. Для этого используется такое понятие, как бэггин. [16].

При бэггинге есть возможность построения огромного количества деревьев, которые минимально коррелируют между собой. При этом ограничивается набор предикторов для каждого дерева – это позволяет минимизировать проблему переобучения. Точность и масштабируемость модели зависит от количества деревьев решений.

Однако, в отличие от деревьев решений, случайный лес не так легко интерпретируется. Например, невозможно показать, что определенное событие состоится в определенное время – известно только, что это заключение большинства деревьев, включенных в лес. Недостаток ясности не позволяет применять данную модель в таких областях, как, например, медицина. Эффективность случайных лесов применима в ситуациях, когда необходима точность, а не интерпретируемость [17].

***K*-ближайшие соседи**

Можно использовать один из простейших методов машинного обучения – *k*-ближайшие соседи (также известный, как *k-NN*). При применении данного метода, значения относят к тому из классов, к которому относится большинство из k его соседних значений. Например, если $k = 5$, значение сравнивается с 5 своими ближайшими объектами-значениями.

При обучении метод запоминает признаки (которые являются векторами) и те классы, которым принадлежат объекты. Когда дело доходит до реальных данных (также называемыми наблюдениями) с неизвестными классами, считается расстояние между вектором наблюдения и уже известными векторами. После этих вычислений происходит выборка k ближайших векторов и наблюдение записывается в класс, принадлежащий большинству среди них. Данный метод также можно использовать и для задач регрессии, в этом случае наблюдениям

будет присваиваться среднее значение ближайших объектов [18].

При использовании данного метода возникают основные проблемы: выбор числа соседей, отсев выбросов, сверхбольшие выборки, выбор метрики.

Главный параметр для данного метода – выбор числа *k*. С одной стороны, чем больше это число, тем информация достовернее, однако, границы между классами становятся все более размытыми. При крайних случаях, когда это число равно 1 или количеству всей выборки, классификация соответственно неустойчива к шумовым выбросам, либо, наоборот, слишком устойчива и становится константой. Хороший результат можно обеспечить при помощи эвристических алгоритмов для выбора числа, к примеру, перекрестная проверка (*cross-validation*) [19].

Среди объектов обучения можно выделить следующие:

- эталоны – очевидные объекты для какого-либо из классов;
- периферийные – окруженные объектами класса, к которому они относятся: удаление их из выборки не отражается на качестве классификации;
- шумовые выбросы – неверно классифицированные объекты: удаление их из выборки повышает качество классификации.

Также, удаление периферийных объектов и шумовых выбросов естественно сокращает объем изначальной выборки, что позволяет уменьшить время для обработки данных. Для выбора же эталонов возможно использование такого алгоритма, как *STOLP*.

В случае сверхбольших выборок (для данного метода более 1000), выборка либо прореживается путем удаления неинформативных данных, как описано выше, либо возможно применение специальных структур данных и индексов, при которых возможен ускоренный поиск соседей (пример: *kd*-деревья).

Самой сложной задачей является проблема выбора метрики (функции близости). При применении метода ближайших соседей редко бывает изначально известна функция расстояния. Например, для числовых данных чаще всего используется Евклидово расстояние. В основном, проблема решается выбором наиболее информативных признаков, которое создает множество разных наборов, для каждого строится своя функция близости и потом методом голосования выбирается одна.

Наивный байесовский классификатор

Наивный байесовский классификатор является одним из наиболее простых классификаторов, использующихся в интеллектуальном анализе данных. Он основан на теореме Байеса, в которой присутствуют строгие (наивные) предположения независимости.

В целом, основной задачей при классификации является восстановление плотностей (функций правдоподобия для каждого класса). Существует несколько подходов: параметрическое восстановление (предполагается, что плотности – гауссовские), непараметрическое восстановление, разделение смеси распределений.

Подход в случае наивного классификатора основывается на том, что плотности распределения для классов уже известны. Алгоритм считается оптимальным из-за минимальной вероятности ошибок [20]. Для наивного классификатора существует дополнительное предположение: объекты описаны комплексом независимых признаков. Независимость признаков ведет к упрощению задачи – оценка нескольких *n* одномерных плотностей проще, чем оценка одной *n*-мерной. Представление классификатора может быть как параметрическим, так и не параметрическим – все зависит от метода восстановления одномерных плотностей.

На этом обзор основных методов контролируемого обучения окончен. В неконтролируемом обучении главной задачей является нахождение закономерностей (выводов) в исходных данных. Для этого используется такой широкий спектр методов, как кластерный анализ. Используя кластеризацию, возможно нахождение взаимосвязи между переменными (наблюдениями). При всем этом выходных параметров, которые являются гипотетическими, при анализе нет. Именно поэтому задача кластеризации отличается от задач классификации – она пытается создать несколько групп данных по только имеющимся входным переменным. Объекты каждой из групп должны быть однородными по какому-либо из признаков и отличаться от других групп (кластеров) [21].

Существует несколько типов кластеризации, широко применяемых в машинном обучении и интеллектуальном анализе данных: иерархическая кластеризация, кластеризация методом *k*-средних, анализ основных компонентов. Также стоит упомянуть о методах кластеризации и ранжирования, используемых в теории графов.

Иерархическая кластеризация

Иерархическая кластеризация – это комплекс алгоритмов кластеризации, которые создают иерархию вложений (разбиений) исходного набора данных. Также в литературе можно встретить такое название иерархических алгоритмов, как алгоритмы таксономии. При визуализации кластерного разбиения данных используются дендрограммы – древовидные структуры, строящиеся по матрицам мер близости между кластерами.

В каждом узле дерева расположены подмножества начальной выборки, на каждом из уровней количество объектов также равно количеству объектов в изначальном наборе.

В общем случае алгоритм иерархической кластеризации описывается так: в отличие от деревьев решений, дерево кластеризации

строится, начиная с листьев и заканчивая корнем. Изначально каждый элемент – это отдельный кластер. По мере работы алгоритма на каждой итерации происходит объединение двух кластеров, находящихся ближе всего друг к другу. Процесс продолжается, пока не будет получено определенное количество кластеров [22].

Для объединения кластеров необходимо вычислять расстояние между ними. На начальном этапе кластеризации используется выбранная мера расстояния. Возможны следующие меры:

- евклидово расстояние – геометрическое расстояние в пространстве, общий тип для большинства задач;
- квадрат евклидова расстояния – используется для возможности установить больший вес наиболее отдаленным объектам;
- расстояние городских кварталов (манхэттэнское расстояние) – средняя разность по координатам, результаты схожи с евклидовым расстоянием, но более устойчиво к выбросам;
- расстояние Чебышева – используется для кластеризации объектов, отличающихся по только одной из характеристик (координате);
- степенное расстояние – используется для изменения веса, который относится к отличающейся характеристике объекта;
- процент несогласия – используется в случае категориальных данных.

На следующих же шагах уже требуются правила объединения кластеров с несколькими объектами. Для этого применяются различные функции, в зависимости от постановки задачи. Примеры таких функций:

- метод одиночной связи (метод ближайшего соседа) – рассматривается каждое значение кластера и ищется ближайшее значение из другого кластера;
- метод полной связи (метод дальнего соседа) – противоположность метода ближайшего соседа – берется максимальное расстояние между значениями двух кластеров;
- метод средней связи – берется среднее расстояние между значениями;
- центроидный метод – расстояние равно расстоянию между центрами кластеров;
- метод Уорда – рассматривается каждая пара кластеров и вычисляется дисперсия при их объединении, минимальная дисперсия является расстоянием [23].

Кластеризация методом k-средних

Самым известным неиерархическим алгоритмом кластеризации является метод *k*-средних. При его использовании пространство данных разбивается на определяемое заранее количество *k* кластеров. Алгоритм в общем виде описывается следующим образом:

1. Определяется число *k* – количество кластеров;
2. В исходном наборе данных выбирается количество объектов, равное *k*. Данные объекты являются изначальными центрами кластеров.
3. Для каждого из объектов выбирается наиболее близкий центр. Особенностью метода является постоянная мера расстояния – евклидово расстояние. Образуются начальные кластеры.
4. Вычисляются центры тяжести для каждого из кластеров, также называемые центроидами. Центроид представляет собой вектор средних значений признаков для всех объектов в кластере.
5. Центр смещается в центроид и становится новым центром кластера.
6. Шаги №3 и №4 повторяются до момента остановки. Итеративный процесс используется для изменения границ кластеров и смещения центров. Расстояние между кластерами увеличивается, а расстояние между элементами внутри кластера постепенно уменьшается [24].

Момент остановки определяется тогда, когда в каждом кластере остаются одни и те же объекты. Об этом также свидетельствует неизменность центроидов и границ кластеров. Для практических задач стабилизация кластеров происходит за 10-20 итераций.

Главные преимущества метода – скорость работы и простота реализации. Главным недостатком же является неопределенность в задании количества кластеров для начала работы алгоритма. Также возможны ситуации, когда объект находится посередине 2 центров, что может привести к постоянному перемещению объекта из одного кластера в другой, так как кластеры должны быть цельными и объект не может быть включен в несколько кластеров одновременно [25].

Одним из способов определения оптимального числа *k* может быть использование графика осыпи (графика Кеттела). Он показывает прямую зависимость внутрикластерного разнообразия от числа *k*. Острый изгиб, называемый изломом, предлагает оптимальное количество кластеров [26].

Анализ основных компонентов

Анализ основных компонентов (также можно встретить название метод главных компонент - *PCA*) – метод нахождения главных переменных (основных компонентов), которые разделяют исходную выборку определенным оптимальным способом. Эти переменные предполагают наибольший разброс для данных.

Обычно переменные для разделения представлены в различных единицах. Для получения комбинаций необходимо провести их стандартизацию. Стандартизация – процесс представления каждой переменной в процентилях, которые, в свою очередь, конвертируют переменные в общую (единую) шкалу. Весь этот процесс позволяет проводить вместо анализа, например, по четырем различным переменным, анализ по меньшему количеству скомбинированных высококоррелирующих переменных. В связи с этим, анализ основных

компонентов относится к методам уменьшения размерности [27].

Выбор количества компонент производится схожим образом, как и в кластеризации методом *k*-средних. Количество определяется графиком осыпи, оптимальное число соответствует положению излома.

Метод главных компонент, несмотря на свою простоту в анализе данных с несколькими переменными, имеет некоторые ограничения. Одно из них – максимизация распределения. Метод допускает, что наиболее полезными являются измерения, дающие больший разброс, что при постановке разных задач не всегда так. Известным примером, показывающим это ограничение, является подсчет блинов в стопке. Если стопка невысокая, метод ошибочно предположит, что главным компонентом будет диаметр блина из-за нахождения в этом измерении большего разброса.

Основная трудность метода – интерпретация компонентов. Необходимы предварительные знания о категориях переменных. Это помогает объяснить, почему именно эти переменные скомбинированы.

Кластеризация и ранжирование графов

Также для кластеризации возможно использование теории графов. Например, для социальной сети из нескольких пользователей можно построить графовое представление, состоящее из узлов (собственно пользователей сети) и ребер (отношений пользователей). Ребра могут иметь свой определенный вес, который показывает силу отношений. Кроме пользователей, графы могут использоваться и для построения схем между другими сущностями, если они связаны каким-либо образом между собой.

Для визуализации таких графов отношений используется силовой алгоритм – узлы, которые не связаны друг с другом, отталкиваются друг от друга, а связанные – притягиваются, учитывая силу связи (степень близости) [28].

Получившийся граф можно прокластеризовать различными методами. Наиболее известным является лувенский метод [29]. Данный метод позволяет подобрать разные конфигурации кластеров для решения следующих задач:

- максимизация количества и силы связей между узлами для одного кластера;
- минимизация связей между узлами разных кластеров.

Степень удовлетворения вышеупомянутым условиям называется модулярностью. Чем выше модулярность, тем больше вероятность того, что кластеры построены наиболее оптимально.

Для получения оптимальной конфигурации метод следует следующему алгоритму:

1. Каждый узел рассматривается как отдельный кластер. На этом этапе количество узлов равно количеству кластеров.
2. Меняется кластерное членство узла, если это улучшает модулярность. В противном случае, узел остается в этом же кластере. Процесс повторяется для каждого из узлов, пока изменения не закончатся.
3. Строится грубая версия, в ней кластеры, полученные на 2 шаге, представляются отдельными узлами. Также происходит объединение предыдущих межкластерных отношений в утолщенные ребра, соответствуя их весу.
4. Повторений шагов 2 и 3 до окончания изменений членства и размера связей.

Результатом является выделение наиболее значимых кластеров. Популярность метода обусловлена простотой реализации и эффективностью, однако, существуют некоторые ограничения.

В ходе слияния значимые, но небольшие по размеру кластеры могут быть поглощены. Для решения этой проблемы возможно использование промежуточной проверки идентифицированных кластеров.

Также процесс принятия решения, какая конфигурация является оптимальной, может быть достаточно трудоемким. Если существует несколько решений с одинаково высокой модулярностью, необходима дополнительная информация.

Помимо кластеризации для анализа графов используется ранжирование – нахождение ведущих узлов, вокруг которых соответственно формируются кластеры. Одним из алгоритмов, который используется для данных задач, является алгоритм *PageRank*. *PageRank* изначально использовался для ранжирования веб-сайтов и значения для каждого из них определялось тремя факторами: число, сила и источник ссылок [30]. Алгоритм достаточно прост в своем использовании, но главным недостатком является необъективность в отношении старых узлов. В связи с этим, необходимо постоянное обновление по мере добавления новых ссылок.

ЗАКЛЮЧЕНИЕ

В качестве предметной области в данной работе выступает информация, получаемая из социальных медиа. Так как чаще всего социальные медиа являются структурой, представляемой в виде графов, в настоящее время мы предполагаем, что наиболее адекватным методом для поставленной задачи будет «Кластеризация и ранжирование графов». По итогам рассмотрения достоинств и недостатков различных методов предиктивного анализа, из методов контролируемого обучения также можно выделить методы случайного леса и *k*-ближайших соседей, как возможные для использования в некоторых дополнительных задачах анализа данных.

ЛИТЕРАТУРА

[1] Прогнозная аналитика: 10 примеров, инструменты и алгоритмы. In-scale: маркетинг для руководителей. URL:

https://iot.ru/promyshlennost/vozmozhnosti-prognoznoy-analitiki-keys-ot-beltel-datanomics (Дата обращения: 01.10.2019).
[2] Предиктивная аналитика: 7 примеров использования в бизнесе. Блог Ingate: SEO, медийная реклама, интернет-маркетинг. URL: https://blog.ingate.ru/detail/7-primerov-ispolzovaniya-prediktivnoy-analitiki-v-biznese/ (Дата обращения: 01.10.2019).
[3] Daniel D. Gutierrez. Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. Technics Publications, 2015. 282 p.
[4] 5 видов регрессии и их свойства – NOP: Nuances of programming. Medium. URL: https://medium.com/nuances-of-programming/5-видов-регрессии-и-их-свойства-f1bb867aebcb (Дата обращения: 03.10.2019).
[5] Е.З. Демиденко. Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981. – 302 с.
[6] Стохастический градиентный спуск. MachineLearning.ru. URL: http://www.machinelearning.ru/wiki/index.php?title=Стохастический_градиентный_спуск (Дата обращения: 03.10.2019).
[7] Множественная регрессия. Портал знаний StatSoft – электронный учебник по статистике. URL: http://statsoft.ru/home/textbook/modules/stmulreg.html (Дата обращения: 03.10.2019).
[8] Модель полиномиальной регрессии. Хабр. URL: https://habr.com/ru/post/414245 (Дата обращения: 04.10.2019).
[9] Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 1. В 2-х кн. М.: Финансы и статистика, 1986. – 366 с.
[10] Классификация и регрессия с помощью деревьев принятия решений. Хабр. URL: https://habr.com/ru/post/116385/ (Дата обращения: 06.10.2019).
[11] Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Электронная книга. URL: https://github.com/ranalytics/data-mining
[12] Как легко понять логистическую регрессию. Хабр. URL: https://habr.com/ru/company/io/blog/265007/. (Дата обращения: 08.10.2019).
[13] Логистическая регрессия и ROC-анализ – математический аппарат. BaseGroupLabs. URL: https://basegroup.ru/community/articles/logistic (Дата обращения: 08.10.2019).
[14] Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М.: МЦМНО, 2013. – 390 с.
[15] Bernhard E. Boser, Isabelle M. Guyon, Vladimir N. Vapnik A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92), page 144-152. Pittsburgh, PA, USA, ACM Press, (July 1992).
[16] Ансамбли моделей: бустинг и бэггинг. DataReview.info – Ваш проводник в мире анализа данных – URL: http://datareview.info/article/ansambli-modeley-busting-i-begging/ (Дата обращения: 11.10.2019).
[17] Анналин Ын, Кеннет Су. Теоретический минимум по Big Data. СПб.: Питер, 2019. – 208 с.
[18] Алгоритм k-ближайших соседей. Data Science – Наука о данных по-русски. URL: http://datascientist.one/k-nearest-neighbors-algorithm/ (Дата обращения: 13.10.2019).
[19] Шалев-Шварц Шай, Бен-Давид Шай. Идеи машинного обучения. М.: ДМК Пресс, 2019. – 432 с.
[20] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989. – 607 с.
[21] Обучение без учителя: 4 метода кластеризации на Python. ProgLib – библиотека программиста. URL: https://proglib.io/p/unsupervised-ml-with-python/ (Дата обращения: 15.10.2019).
[22] Иерархическая кластеризация – викиконспекты. Университет ИТМО URL: http://neerc.ifmo.ru/wiki/index.php?title=Иерархическая_кластеризация (Дата обращения: 16.10.2019).
[23] Иерархическая кластеризация. Создание децентрализованных приложений на блокчейн. Помощь в программировании искусственного интеллекта. URL: https://craftappmobile.com/иерархическая-кластеризация (Дата обращения: 16.10.2019).
[24] Hartigan J. A., Wong, M. A. Algorithm AS136: a k-means clustering algorithm. Applied Statistics. 1979. Vol. 28. pp. 100-108.
[25] Алгоритм k средних (k-means). Создание децентрализованных приложений на блокчейн. Помощь в программировании искусственного интеллекта. URL: http://algowiki-project.org/ru/Алгоритм_k_средних_(k-means) (Дата обращения: 17.10.2019).
[26] Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка, М. С. Олдендерфер, Р. К. Блэшфилд. Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. – 215 с.
[27] Иберла К. Факторный анализ. М.: Статистика, 1980. – 398 с.
[28] Demetrescu C., Finocchi I., Stasko J.T. Specifying Algorithm Visualizations: Interesting Events or State Mapping? In Proc. of Dagstuhl Seminar on Software Visualization – Lect. Notes in Comput. Sci. 2001. P. 16–30.
[29] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008, 2008.
[30] Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, Stanford University (1998).

Иван Сергеевич Калытюк, аспирант кафедры автоматики НГТУ. Основное направление научных исследований: разработка и исследование систем сбора и анализа данных.

E-mail: ivankalytyuk@yandex.ru

Новосибирск, просп. Карла Маркса, д.20, НГТУ

Галина Александровна Французова, д. т. н., профессор кафедры автоматики НГТУ. Основное направление научных исследований: синтез систем экстремального регулирования.

E-mail: frants@ac.cs.nstu.ru
Новосибирск, просп. Карла Маркса, д.20, НГТУ

Андрей Васильевич Гунько, к. т. н., доцент кафедры автоматики НГТУ. Основное направление научных исследований: разработка автоматизированных систем сбора и обработки результатов.
E-mail: gun@ait.cs.nstu.ru
Новосибирск, просп. Карла Маркса, д.20, НГТУ

Статья поступила 20 сентября 2019 г.

On the Choice of Methods of Predictive Analysis of Social Media Data

I. S. Kalytyuk, G. A. Frantsuzova, A.V. Gunko
Novosibirsk State Technical University, Novosibirsk, Russia

Abstract. This article discusses the issue of the choice of predictive analysis methods in objectives, related to the analysis of data, obtained from social media. With the development of the Internet and social media, it has become easier to access and distribute information, because the network users themselves are both creators and recipients of various information. The main type of social media is social networks. Instagram, Facebook, VK, Instagram, YouTube, Twitter, Odnoklassniki, WhatsApp and Telegram messengers are among the most well-known ones. The most important functions of social media are to influence the perception, attitude, and final behavior of consumers. Predictive analytics is based on automatic search for relationships, anomalies, and patterns between various factors. This method of analyzing social media data is at the stage of its development. Hence, the relevance of the study of predictive analytics methods on the example of the analysis of social media data extracted from open sources is obvious. The results of predictive analytics can be of practical value in the following cases: creating recommendation services for interested users in a particular product/service, predicting accidents/failures, making optimal decisions. The scientific novelty is the assessment of the applicability of predictive analytics methods in the field of social media data analysis. Possible methods of controlled and uncontrolled learning, such as regression, classification, clustering, are considered. Regression methods such as multiple linear regression, polynomial regression, and regression trees are described in detail. The following known methods are distinguished from the classification methods: logistic regression, classification trees, supporting vector machines, random forests, k-nearest neighbors, naive Bayesian classifier. In the field of clustering, attention is paid to the methods of hierarchical clustering, k-means clustering, principal component analysis, clustering and ranking methods used in graph theory. The algorithms that are most adequate for this subject area, which can lead to new information useful for developers and users, are highlighted.

.

Keywords: predictive analysis, social media, regression, classification, clustering.

REFERENCES

[1] Predictive Analytics: 10 examples, tools and algorithms. Available at: https://iot.ru/promyshlennost/vozmozhnosti-prognoznoy-analitiki-keys-ot-beltel-datanomics (accessed 1 October 2019).

[2] Predictive Analytics: 7 business use cases. Available at: https://blog.ingate.ru/detail/7-primerov-ispolzovaniya-prediktivnoy-analitiki-v-biznese/ (accessed 1 October 2019).

[3] Daniel D. Gutierrez. Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. Technics Publications, 2015. 282 p.

[4] 5 types of regression and their properties - NOP: Nuances of programming Available at: https://medium.com/nuances-of-programming/5-видов-регрессии-и-их-свойства-f1bb867aebcb (accessed 3 October 2019).

[5] E.Z. Demidenko. Linear and nonlinear regression. Moscow, Finance and statistics, 1981. 302 p.

[6] Stochastic gradient descent. Available at: http://www.machinelearning.ru/wiki/index.php?title=Стохастический_градиентный_спуск (accessed 3 October 2019).

[7] Multiple regression. Available at: http://statsoft.ru/home/textbook/modules/stmulreg.html (accessed 3 October 2019).

[8] Polynomial regression model. Available at: https://habr.com/ru/post/414245 (accessed 4 October 2019).

[9] Dreyper N.. Smit G. Applied regression analysis. Book 1. Moscow, Finance and statistics, 1986. 366 p.

[10] Classification and regression using decision trees. Available at: http://habr.com/ru/post/116385/ (accessed 6 October 2019).

[11] Shitikov V.K. , Mastickij S.E. (2017) Classification, regression and other Data Mining algorithms using R. Elecronic book, available at: https://github.com/ranalytics/data-mining

[12] How easy it is to understand logistic regression. Available at: https://habr.com/ru/company/io/blog/265007/ (accessed 8 October 2019).

[13] Logistic regression and ROC analysis-mathematical apparatus. Available at: https://basegroup.ru/community/articles/logistic (accessed 8 October 2019).

[14] V'yugin V.V. Mathematical foundations of the theory of machine learning and forecasting. Moscow, MCMNO, 2013. 390 p.

[15] Bernhard E. Boser, Isabelle M. Guyon, Vladimir N. Vapnik A training algorithm for optimal margin classifiers // Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92), page

144-152. Pittsburgh, PA, USA, ACM Press, (July 1992)
[16] Ensembles of models: boosting and bagging Available at: http://datareview.info/article/ansambli-modeley-busting-i-begging/ (accessed 11 October 2019).
[17] Annalin Yn, Kennet Su. Theoretical minimum for Big Data. St. Petersburg, Piter, 2019. 208 p.
[18] K-nearest neighbor algorithm. Available at: http://datascientist.one/k-nearest-neighbors-algorithm/ (accessed 13 October 2019).
[19] Shaj, Ben-David Shaj. The idea of machine learning. Moscow, DMK Presss, 2019. 432 p.
[20] Ajvazyan S. A., Buhshtaber V. M., Enyukov I. S., Meshalkin L. D. Applied statistics: classification and dimension reduction. Moscow, Finance and statistics, 1989. 607 p.
[21] Learning without a teacher: 4 clustering methods in Python. Available at: https://proglib.io/p/unsupervised-ml-with-python/ (accessed 15 October 2019).
[22] Hierarchical clustering. Available at: http://neerc.ifmo.ru/wiki/index.php?title=Иерархическая_кластеризация (accessed 16 October 2019).
[23] Hierarchical clustering. Available at: https://craftappmobile.com/иерархическая-кластеризация (accessed 16 October 2019).
[24] Hartigan J. A., Wong, M. A. Algorithm AS136: a k-means clustering algorithm. Applied Statistics. 1979. Vol. 28. pp. 100-108.
[25] K-means algorithm. Available at: http://algowiki-project.org/ru/Алгоритм_k_средних_(k-means). (accessed 17 October 2019).
[26] Dzh.-O. Kim, CH. U. M'yuller, U. R. Klekka, M. S. Oldenderfer, R. K. Bleshfild. Factor, discriminant and cluster analysis. Moscow, Finance and statistics, 1989. 215 p.
[27] Iberla K. Factor analysis. Moscow, Statictics, 1980. 398 p.
[28] Demetrescu C., Finocchi I., Stasko J.T. Specifying Algorithm Visualizations: Interesting Events or State Mapping? // In Proc. of Dagstuhl Seminar on Software Visualization – Lect. Notes in Comput. Sci. – 2001. – P. 16–30.
[29] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008, 2008.
[30] Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, Stanford University (1998).

Ivan Sergeevich Kalityuk, PhD student of the Department of Automation NSTU. The main direction of scientific research: development and research of data collection and analysis systems.

E-mail: ivankalytyuk@yandex.ru

Novosibirsk, prosp. Karl Marx, 20

Galina Aleksandrovna Frantsuzova, Doctor of Technical Sciences, Professor, Department of Automation, NSTU. The main direction of scientific research: the synthesis of systems of extreme regulation.

E-mail: frants@ac.cs.nstu.ru

Novosibirsk, prosp. Karl Marx, 20

Andrei Vasilievich Gunko, Ph.D., associate professor of the Department of Automation NSTU. The main direction of scientific research: the development of automated systems for collecting and processing results.

E-mail: gun@ait.cs.nstu.ru

Novosibirsk, prosp. Karl Marx, 20

The paper has been received on 20/09/2019.