Интеллектуальные методы поиска следов мошенничества в больших данных

Станислав ЗВЕЖИНСКИЙ, д.т.н., профессор МТУСИ
Владимир ИВАНОВ, д.т.н., профессор, эксперт

ОБЩИЕ ПОЛОЖЕНИЯ

Мошенничество с использованием цифровых данных стало настоящим бичом информационного сообщества: подделка документов, манипулирование на электронных торгах, недостоверная клиентская информация, несанкционированная идентификация личности и пр. [1,2]. Такая информация, которую будем в общем случае называть аномалией, как правило, «размазана» или «маскируется» в огромном объеме т.н. «больших данных» (БД), что не позволяет ее «вручную» обнаруживать и анализировать. Современные методы искусственного интеллекта (ИИ) способны, с той или иной степенью достоверности, решить эту проблему. Бурное развитие методов ИИ в последние 10 лет, начало которым положено более века назад, обусловлено качественным скачком вычислительной способности современных компьютеров.

Большие данные (англ. Big Data, БД) – в общем случае неструктурированные и многообразные данные огромных объемов. Обработка БД на современном этапе осуществляется преимущественно с использованием методов ИИ. Для этого широко используются парсинг (англ. parsing) – синтаксический анализ разнородной информации с приведением ее к единой форме. Известно более 20-ти алгоритмов парсинга данных, к наиболее известным относятся LogSig, LKE, MoLFI, LFA, SHISO, LogMine, LtnMa, Spell, AEL, IPLoM. Качество парсинга во-многом определяет эффективность работы с БД. Поиск и анализ аномалий в БД может осуществляться многочисленными классами методов [3], из которых выделяется:

- машинное обучение (англ. Machine learning, МО) – разнообразный математический инструментарий для решения интеллектуальных задач;
- интеллектуальный анализ данных (англ. Data mining, ИАД) – совокупность методов обнаружения в данных ранее неизвестных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений;
- классический статистический анализ (корреляционный, регрессионный, дисперсионный и пр.), основанный на классических науках: информатике, теории вероятностей и пр.;
- искусственные нейронные сети (ИНС);
- методы оптимизации, в том числе генетические алгоритмы;
- кластерный анализ;
- распознавание образов;
- статистический анализ (классический) и др.

Методы ИИ могут относиться к разным классам, имеются их «пересечения» [3-5]. На рис. 1 показана взаимосвязь БД и основных классов методов по их обработке.

Рисунок 1.
Взаимосвязь больших данных и групп методов по их обработке

Как видно из рис. 1, ИАД и МО – два «пересекающихся» класса области ИИ. Первые характеризуют процесс выявления закономерностей и получения новой, неизвестной информации из исходного массива необработанных данных; вторые – это, по сути, методы построения алгоритмов анализа, способных обучаться.

ИАД подразумевает процесс обнаружения в «сырых» данных знаний, ранее неизвестных, практически полезных и доступных к интерпретации, описывающих новые связи, предсказывающие значения одних признаков на основе других и т.д. Знания должны быть представлены в понятном для пользователя нематематическом виде, например, в виде логической конструкции «если ..., то ...». К классу ИАД относят некоторые методы, не включенные в МО: генетические алгоритмы, эволюционное программирование, нечеткая логика [3]. К ИАД нередко относят и классический статистический анализ (корреляционный, регрессионный, факторный анализ), предполагающий априорные представления об анализируемых данных, а также процесс «очистки» БД и их подготовки к последующему анализу.

Наиболее распространенные методы и алгоритмы ИАД – это ИНС, деревья решений, алгоритмы кластеризации и обнаружения ассоциативных связей между событиями, эти же методы (алгоритмы) зачастую относят и к области МО, чьей основной проблематикой является самостоятельное получение знаний некой интеллектуальной системой (ИС) в процессе ее работы и обучения (это направление было центральным с самого начала развития ИИ). Под интеллектуальной системой (англ. intelligent system, ИС) понимается информационно-вычислительный комплекс, способный без участия лица, принимающего решение (ЛПР), решать творческие задачи из конкретной

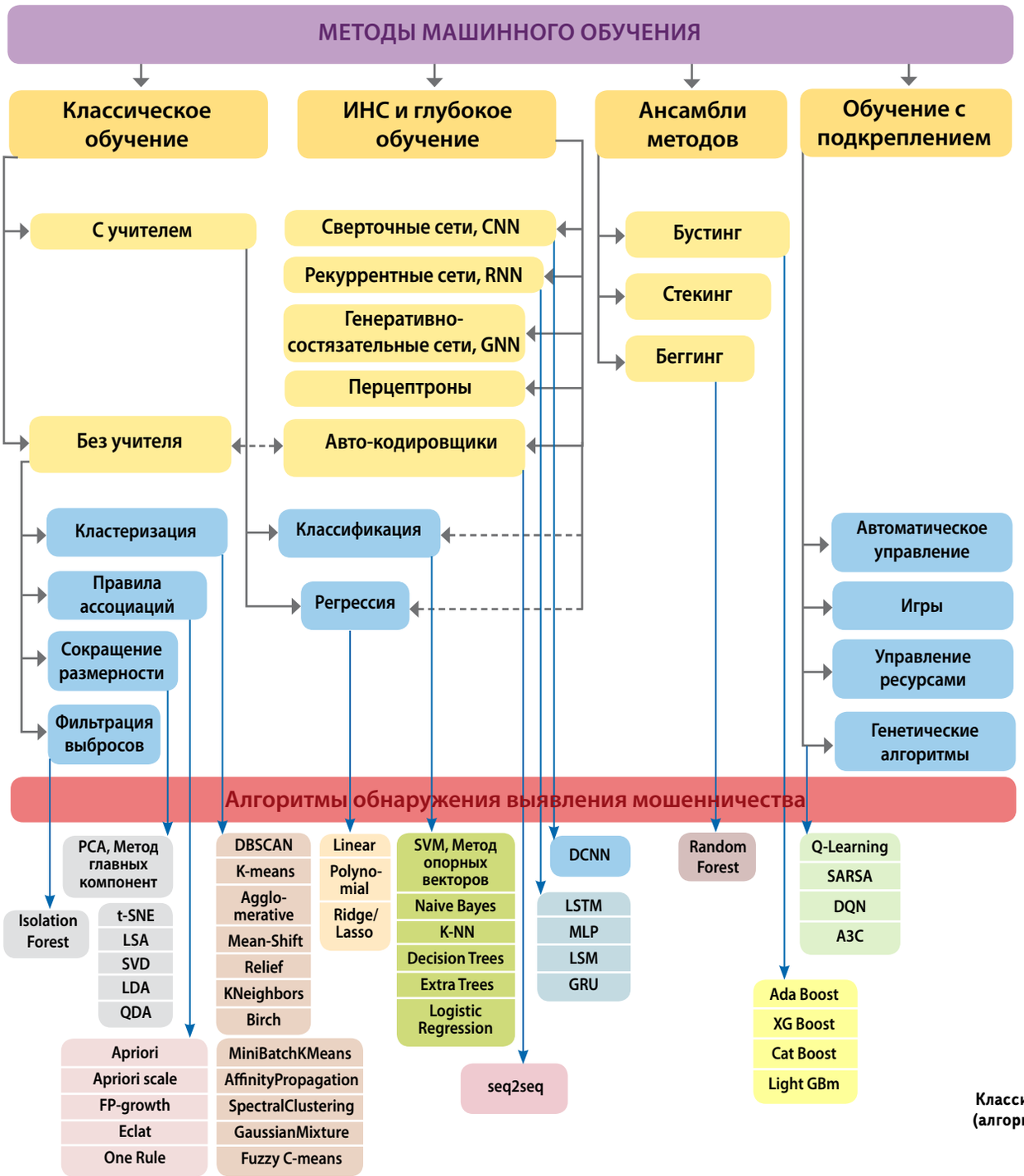

Рисунок 2. Классификация методов (алгоритмов) машинного обучения

предметной области, знания о которой хранятся в ее памяти. ИС включает три основных компонента: базу знаний, механизм принятия решений и интеллектуальный интерфейс.

Из анализа многочисленных работ по ИИ следует, что не существует единственного оптимального метода ИАД или МО по поиску аномалий в БД, все зависит от конкретных условий объекта и предмета исследований. Под последним будем понимать реализуемые с помощью ИС интеллектуальные методы выявления аномалий в БД, связанных со следами мошенничества.

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Машинное обучение – это методы ИИ, характеризующиеся не прямым решением задач (как при классическом статистическом анализе), а обучением в процессе решения [3]. Различаются два типа обучения: 1) обучение по прецедентам (в том числе самообучение), основанное на выявлении эмпирических закономерностей в исходных данных; 2) дедуктивное обучение, которое предполагает формализацию знаний экспертов в виде построения базы знаний (БЗ). Дедуктивное принято относить к области экспертных систем, которые при анализе БД практически не используются, поскольку имеют низкую точность. В этом смысле термины «машинное обучение» и «обучение по прецедентам» можно считать тождественными. В МО выделяются 4 основных группы (направления): 1) классическое обучение; 2) обучение с подкреплением; 3) ансамбли; 4) ИНС и глубокое обучение. В свою очередь, классическое подразделяется на обучение без учителя (т.н. «индуктивная машина вывода») и обучение с учителем. На рис. 2 представлена одна из возможных классификаций методов МО, не претендующая на полную

корректность, поскольку в литературе те или иные методы приписываются к разным подгруппам.

КЛАССИЧЕСКОЕ МАШИННОЕ ОБУЧЕНИЕ

Оно строится на классических статистических алгоритмах принятия решений на основе данных, где выявляются закономерности. Такое обучение подразделяется на две категории: 1) обучение с учителем, когда обучающая выборка данных «размечена»; 2) обучение без учителя (в том числе самообучение), когда обучающая выборка не «размечена», т.е. нет целевой переменной (признака). В нашем случае разметка на два класса подразумевает наличие аномалии (условно логическая «1») или ее отсутствие (логический «0»).

ОБУЧЕНИЕ С УЧИТЕЛЕМ

На рис. 3 показана схема классификации с учителем; решаемые задачи делятся на две группы: 1) классификация – предсказание категории объекта («1»/«0»); 2) регрессионный анализ – предсказание чисел (на прямой).

КЛАССИФИКАЦИЯ. Ее суть заключается в разделении объектов по заранее известному признаку. Строится модель для прогнозирования категориальных меток неизвестных объектов, чтобы различать их по классам; эти категориальные метки предварительно заданы, дискретны и не упорядочены. При обучении алгоритму подается заранее размеченный набор признаков, в нем ищутся взаимосвязи, которые используются для дальнейшего определения классов в неразмеченных данных. Классификация используется в таких областях, как спам-фильтры, выявление подозрительных банковских транзакций. Популярные алгоритмы классификации: 1) наивный байесовский классификатор (Naive Bayes, NB) - семейство; 2) деревья принятия решений (Decision Tree, DT) - семейство; 3) K-ближайших соседей (K-NN); 4) опорных векторов (Support Vector Machine, SVM); 5) логистическая регрессия (Logistic Regression, LR).

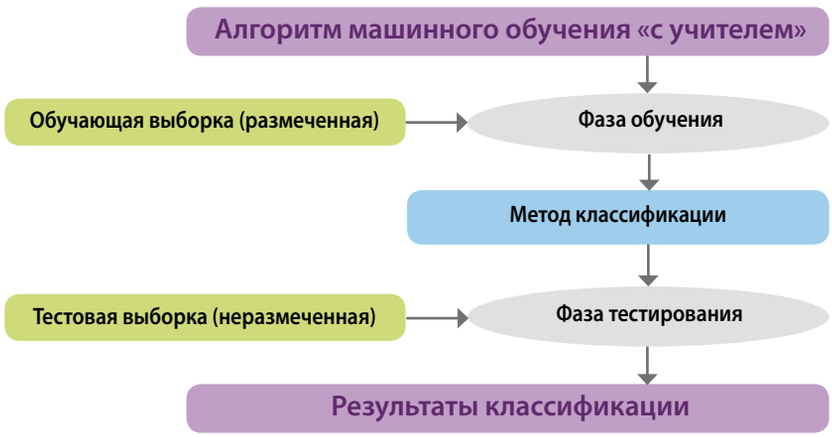

Рисунок 3.
Общая схема классификации методами МО «с учителем»

Наивный байесовский классификатор (NB) – достаточно простой вероятностный классификатор, основанный на теореме Байеса со строгими (наивными) предположениями о независимости всех n признаков, то есть когда значение одного параметра не оказывает влияния на другой [6,7]. Предположение о независимости существенно упрощает задачу, т.к. оценить n одномерных признаков гораздо легче, чем один n-мерный. Алгоритм вычисляет, сколько раз определённые признаки встречаются в каждом классе, и на этом основании принимается решение. Предположение о независимости признаков редко выполняется на практике, что приводит в большинстве задач к недостаточному качеству решений. С другой стороны, NB

работает существенно быстрее многих других и известен как «достойный» классификатор, однако не ищет взаимосвязи между признаками и поэтому в сложных задачах не используется. Gaussian naive bayes (GNB) реализует наивный алгоритм Байеса там, где вероятностные характеристики объектов предполагаются гауссовыми.

Деревья решений (DecisionTreeClassifier, DTC) – семейство алгоритмов, представляющие правила в последовательной структуре, где для каждого объекта есть узел, дающий единственное решение. Правило в данном случае – это логическая конструкция вида «если ... то...». Структура дерева принятия решений состоит из элементов 2-х типов: узлов и листьев. В процессе обучения алгоритм генерирует правила, по которым отделяются объекты разных классов друг от друга. Эти правила размещаются в узлах дерева так, чтобы пройдя от корня к листьям (лист – итоговый класс), определился класс объекта [5, 8]. В простейшем случае от узла отходят две ветви («соответствует / не соответствует» объект признаку, указанному в узле, – «да / нет»). В целом, деревья решений просты в интерпретации результата. Существует множество алгоритмов, реализующих деревья решений - NewId, ITrule, CHAID, CN2 и т.д. Главным образом используются: 1) CART (Classification and Regression Tree) – алгоритм для построения бинарного дерева; 2) C4.5 – алгоритм для построения дерева решений, у которого число потомков узла не ограничено. Большинство алгоритмов «деревьев» являются «жадными», т.е. если один раз выбран признак и по нему произведено разбиение, то алгоритм не может вернуться назад и выбрать другой признак, который мог бы дать лучшее решение.

K-ближайших соседей (k-NearestNeighbors, K-NN) – алгоритм классификации (и регрессии), при котором неизвестный объект присваивается тому классу, который является наиболее распространённым среди K соседей, классы которых уже известны [4, 5]. Свойство алгоритма – только запоминать обучающую выборку, вычисления начинаются на этапе тестирования. Это легко интерпретируемый и изученный алгоритм, суть которого проста: какой класс преобладает вокруг объекта, таков и сам объект; если метрика расстояния выбрана удачно, то схожие объекты «ложатся» рядом. Алгоритм вкратце таков: 1) посчитать расстояние до каждого объекта в обучающей выборке; 2) отобрать k объектов обучающей выборки, расстояние до которых минимально; 3) искомый объект принадлежит классу с преобладающим числом объектов. Для метода K-NN существуют теоремы, утверждающие, что на «бесконечных» выборках это – оптимальный метод классификации. Некоторые исследователи [41] считают его теоретически идеальным алгоритмом, применимость которого ограничена вычислительными возможностями и т.н. «проклятием размерностей». Качество классификации зависит от: 1) числа соседей – настраиваемый (подбираемый) параметр; 2) метрики расстояния между объектами – в (типично евклидово); 3) весов соседей (равные, обратно пропорциональные расстоянию и пр.). Недостаток метода в том, что он не создает каких-либо моделей или правил, обобщающих предыдущий опыт, а лишь основывается на известных объектах.

Метод опорных векторов [5, 9] служит для бинарной классификации, разделения объектов (событий) на два изначально известных класса-множества, например, событий, связанных / не связанных с аномалиями. Объект – это вектор в n-мерном пространстве, где координаты вектора описывают отдельные его атрибуты (признаки). Идея метода – поиск гиперплоскости (или классифицирующей функции), разделяющей векторы на два класса с максимальным зазором в этом пространстве. Чем больше расстояние между крайними объектами, находящимися на границах классов (т.е. опорными векторами), тем меньше ошибка классификатора.

Логистическая регрессия [10] применяется для прогнозирования вероятности возникновения события по значениям множества признаков. Для этого вводится т.н. зависимая переменная, принимающая одно из двух значений «1» / «0» (событие произошло / не произошло), и множество независимых переменных (признаков, предикторов), на основе значений которых вычисляется вероятность принятия значения зависимой переменной. Регрессия позволяет оценить апостериорные вероятности принадлежности объектов двум классам. Вообще регрессивный анализ – это статистическая методология, используемая для выявления взаимосвязи между несколькими независимыми переменными на входе и одной зависимой переменной на выходе. Он используется, чтобы в рядах числовых примеров входа / выхода выявить непрерывную функцию, на основании которой спрогнозировать выход – полезное число. Задачи регрессии и классификации схожи, только вместо класса объекта алгоритм предсказывает число.

Регрессионный анализ. Методы регрессионного анализа в МО используются для обнаружения случаев мошенничества с кредитными картами и корпоративного мошенничества. Популярными алгоритмами являются: линейная и полиномиальная регрессия. Одномерная или линейная регрессия – метод, используемый для моделирования взаимосвязи между одной независимой входной переменной и одной

зависимой выходной. Множественная линейная регрессия создаёт модель взаимосвязи между несколькими входными переменными и выходной зависимой переменной. Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных. Полиномиальная регрессия – когда в алгоритме моделируется не линейное предсказание, а полиномиальная кривая.

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ (ИЛИ САМООБУЧЕНИЕ)

Это группа методов МО, при котором алгоритм спонтанно обучается выполнять поставленную задачу без вмешательства со стороны ЛПР, самостоятельно выявляя закономерности без опоры на размеченные данные. Решаемые задачи обучения без учителя можно разделить на 4 типа: 1) кластеризация – разделение объектов по схожести по разным, в общем случае, неизвестным классам; 2) правила ассоциаций – выявление последовательностей; 3) сокращение размерности – нахождения зависимостей и переход к меньшему числу признаков; 4) фильтрация выбросов – обнаружение нетипичных объектов.

Кластеризация – разбиение заданной выборки объектов на непересекающиеся подмножества - кластеры так, чтобы каждый состоял из схожих, а объекты разных кластеров существенно отличались [11]. Кластеризация – это неконтролируемая классификация без заранее известных классов, и применяется, например, для сегментации рынка (лояльности покупателей). Популярными алгоритмами кластеризации являются метод K-средних, DBSCAN, Mean-Shift, ИНС Кохонена (то есть «пересечение» групп методов по рис. 1).

Метод K-средних наиболее распространен при кластеризации. Его основная идея заключается в том, что начальные кластеры и их «центры» организуются случайным образом. Затем на каждой итерации заново пересчитываются «центры масс» для всех кластеров, полученных на предыдущем шаге, исходя из расстояния до ближайших объектов, до тех пор, пока вычисленные центры не перестанут смещаться [11, 12]. Недостатком алгоритма является явное задание количества кластеров, на которые требуется разбить множество объектов, а также то, что форма кластеров представляет собой окружность. Алгоритм DBSCAN основан на анализе плотности точек (данных) в некотором пространстве, группируя вместе тесно расположенные точки, объединяя их в кластеры, и помечая как выбросы те, которые находятся «одиноко» в областях с малой плотностью. Алгоритм может находить кластеры произвольной формы, является одним из наиболее часто используемых и упоминаемых в научной литературе [13].

Поиск ассоциативных правил – это методы обучения, позволяющие находить взаимосвязи между переменными, используются, например, для анализа паттернов поведения на веб-сайтах. Популярные алгоритмы: Apriori, FPG, Eclat. Apriori – алгоритм МО для изучения ассоциативных правил в реляционных базах данных, которые должны быть преобразованы к нормализованному бинарному виду (0 / 1) [14]. Алгоритм основывается на идентификации часто встречающихся отдельных элементов в БД и их объединении в наборы до тех пор, пока они станут появляться в базе достаточно часто. Он работает в 2 этапа. На первом находятся часто встречающиеся наборы элементов, на втором из них извлекаются правила. Выявленные частые наборы элементов используются для определения правил ассоциации, которые

подчеркивают общие тенденции в базе. Основной недостаток алгоритма – он вычислительно «медленный» вследствие многократного сканирования базы (столько раз, сколько элементов содержит самый длинный набор).

FPG (Frequent Pattern-Growth) базируется на двухэтапном построении специальной структуры данных – FP-дереве для вычисления популярных наборов элементов [15, 16]. При первом проходе алгоритм подсчитывает встречаемость объектов в наборах и запоминает их в таблице заголовков. При втором проходе алгоритм строит структуру FP-дерева путём вставки наборов часто встречающихся объектов.

Сокращение размерности – проблема заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом существенной информации об объектах выборки. Сокращение размерности используется для поиска похожих документов, риск-менеджмента и пр. Популярные алгоритмы: метод главных компонент, сингулярное разложение (SVD), латентное размещение Дирихле (LDA), латентно-семантический анализ (LSA, pLSA, GLSA).

Метод главных компонент (Principal Components Analysis, PCA) – один из важнейших и изученных способов уменьшить размерность данных, потеряв минимум информации [17,18]. Математически он представляет ортогональное линейное преобразование, отображающее данные из исходного пространства признаков в новое пространство меньшей размерности. Иногда метод PCA называют преобразованиями Кархунена-Лоэва или Хотеллинга. PCA применим всегда и для любых статистических данных, а не только для распределений, близких к нормальным, однако он не всегда эффективно снижает размерность при заданных ограничениях на точность. В таком случае требуется несколько (а не одна) компонент, или вообще не достигается приемлемая точность. PCA – наиболее популярный метод сокращения размерности в таких областях, как распознавание образов, поиск аномалий, компьютерное зрение, подавление шума на изображениях; психодиагностика.

Латентно-семантический анализ – это метод обработки информации на естественном языке, анализирующий взаимосвязь между библиотекой документов и терминами в них встречающимися, и выявляющий характерные факторы (тематики), присущие всем документам и терминам [21]. Латентный семантический анализ основан на статистической оценке сходства слов по их значению. Данный алгоритм подходит для определения тематик текстов и поиска похожих документов.

Фильтрация выбросов. Решаемая задача – обнаружение в обучающей выборке небольшого числа нетипичных объектов - аномалий. В некоторых приложениях такой поиск является самоцелью, например, обнаружение мошенничества; в других случаях аномалии являются следствием ошибок в данных или неточности модели, то есть шумом, который должен быть удален из выборки. Фильтрация выбросов применяется при обнаружении вторжений и мошенничества. Наиболее популярным алгоритмом является Isolation Forest и его производные. **Isolation Forest (изолирующий лес)** – алгоритм обнаружения аномалий, основанный на принципе МонтеКарло (случайности), состоящий из деревьев, каждое из которых строится до исчерпания выборки, для построения ветвления в дереве выбираются случайные признак и расщепление. Для каждого объекта мера нормальности – среднее арифметическое глубин листьев, в которые он попал (изолировался). Алгоритм фактически строит лес из комбинации решающих деревьев, где аномальные точки расположены вблизи корней деревьев. Осуществляется случайное разбиение пространства признаков, так что изолированные (аномальные) точки (объекты) отсекаются от нормальных, кластеризованных данных; окончательный результат усредняется по нескольким запускам [20, 21]. Алгоритм работает так, что выбросы с отличиями в общих статистических характеристиках первостепенно будут попадать в листья (на небольшой глубине дерева), которые там легче «изолировать». Алгоритм хорошо распознает именно выбросы. Алгоритм обладает рядом существенных преимуществ: 1) распознаёт аномалии различных видов, как изолированные точки с низкой локальной плотностью, так и кластеры аномалий малых размеров; 2) его сложность относительно невелика; 3) существенных затрат по памяти не требуется; 4) отсутствуют параметры, требующие подбора; 5) инвариантен к масштабированию признаков; 6) задания метрики или другой априорной информации о типе данных не требуется. В силу этого данный алгоритм, не требующий никакой априорной информации, получил широкое распространение в области МО, во многих случаях превосходя другие в сравнительных исследованиях [24].

Окончание – в следующем номере.

ЛИТЕРАТУРА

1. Иванов В., Звежинский С. Проблемы фальсификации фото- и видеоматериалов на современном этапе развития цифровизации // Технологии безопасности. – 2021.- №1.
2. Фатьянов А.А. Большие данные в цифровой экономике: ценность и правовые вызовы // Экономика. Право. Общество. – 2018. - №4(16). – С.37-40.
3. https://ru.wikipedia.org/wiki/Большие_данные; / Машинное_обучение; / Data_mining.
4. http://www.machinelearning.ru/wiki/index.php?title=MachineLearning.
5. НОУ ИНТУИТ: Лекция: Введение в машинное обучение (https://www.intuit.ru/studies/courses/10621/1105/lecture/17981).
6. Zhang H. The optimality of naive Bayes. AA. 2004. T.1. №2. p.3-8.
7. Rish I. et al. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001. T.3. №22. p.41-46.
8. Swain P.H., Hauska H. The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics. 1977. T.15. №3. p.142-147.
9. https://ru.wikipedia.org/wiki/Метод_опорных_векторов.
10. http://www.machinelearning.ru/wiki/index.php?title=Логистическая_регрессия.
11. http://www.machinelearning.ru/wiki/index.php?title=Кластеризация.
12. Likas A., Vlassis N., Verbeek J. The global k-means clustering algorithm. Pattern recognition. 2003. T.36. №2. p.451-461.
13. https://ru.wikipedia.org/wiki/DBSCAN.
14. https://basegroup.ru/community/articles/apriori.
15. Borgelt C. An Implementation of the FP-growth algorithm. Proc. 1-st Int. workshop on open source DM: Frequent pattern mining implementations. 2005. p.1-5.
16. https://basegroup.ru/community/articles/fpg.
17. http://www.machinelearning.ru/wiki/index.php?title=Метод_главных_компон.
18. http://data4.ru/pca.
19. Landauer T.K. et al. (ed.). Handbook of latent semantic analysis. Psychology Press. 2013.
20. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest. 8-th IEEE Int. conf. on Data Mining, 2008. IEEE. 2008. p.413-422.
21. Sun L. et al. Detecting anomalous user behavior using an extended isolation forest algorithm: an enterprise case study. arXiv preprint:1609.06676. 2016.
22. Иванов С.М. Методы детектирования аномалий: Курсовая работа. – М.: МГУ, 2017.
23. https://ru.wikipedia.org/wiki/Q-обучение.
24. De Jong K. Genetic-algorithm-based learning. Machine learning. Morgan Kaufmann. 1990. p.611-638.