УДК 004.82

ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ В МОДЕЛЯХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

2020 О. Ю. Лавлинская, В. И. Киселева

Воронежский институт высоких технологий (г. Воронеж, Россия)

В статье дается аналитический обзор таких подходов, как Data Mining, KDID. Рассматриваются тенденции и перспективы Data Mining. Уделяется внимание вопросу подготовки исходных данных в моделях интеллектуального анализа, предлагается авторский подход к подготовке исходных данных, основанный на расчете глобальных и локальных значений среднего по классификационной выборке данных.

Ключевые слова: Data Mining, Big Data, обработка данных, визуализация решений, деревья решений, Нейронные сети, KDID, проблема подготовки данных

Увеличивающиеся темпы роста данных привели к развитию таких подходов, как Data Mining [1] (KDID – обнаружение знаний в данных или интеллектуальный анализ данных) и Big Data (серия подходов для анализа сверхбольших массивов данных на вычислительных системах) [2].

Для работы с большими массивами данных применяются средства, используемые на вычислительных системах при помощи методов параллельного программирования. В настоящее врем данная проблема активно изучается научными кругами и крупными компаниями, которые в свою очередь, предоставляют разработанные ими продуты, используемые не только крупным бизнесом, но и представителями среднего и малого. Наиболее востребованной задачей является задача анализа научных данных [3], так как распространение «облачных» ресурсов позволило использовать увеличенные ресурсы, ранее недоступные небольшим компаниям.

Главной задачей, в данный момент, являются разработка, внедрение и применение методов Data Mining в активно развивающихся областях жизни, требующих инструментов для решения поставленных задач.

Наращивание мощностей вычислительных систем решает только часть проблем, возникающих при обработке данных, другими задачами являются эффективное распараллеливание вычислений без особого ущерба для производительности и адаптация алгоритмов под архитектуру вычислительной системы.

Data Mining переводится как “добыча” или “раскопка данных”. Нередко рядом с Data Mining встречаются слова “обнаружение знаний в базах данных” (knowledge discovery in databases) и “интеллектуальный анализ данных”. Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных. Основу методов data mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики. К методам data mining нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов, анализ выживаемости, анализ связей). Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями data mining (обнаружение ранее неизвестных нетривиальных и практически полезных знаний). Одно из важнейших назначений методов data mining состоит в наглядном представлении результатов вычислений (визуализация), что позволяет использовать инструментарий data mining людьми, не имеющими специальной математической подготовки,

Лавлинская Оксана Юрьевна – Воронежский институт высоких технологий, доцент, канд. техн. наук, lavlin2010@yandex.

Киселева Валерия Игоревна – Воронежский институт высоких технологий, студент магистратуры, kiseliova.valeri7@yandex.ru.

что и стало причиной бурного развития метода, но и проявило свою специфику к обработке информации:

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки «сырых» данных должны быть просты в использовании.

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборке и виде распределений значений анализируемых показателей.

Рассмотрим задачи, решаемые Data Mining:

1. Классификация – отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов.
2. Кластеризация – разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга.
3. Сокращение описания – для визуализации данных, упрощения счета и интерпретации, сжатия объемов собираемой и хранимой информации.
4. Ассоциация – поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя».
5. Прогнозирование – нахождение будущих состояний объекта на основании предыдущих состояний (исторических данных).
6. Анализ отклонений – например, выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы.
7. Визуализация данных.

В связи с этим, Data Mining используется в широком кругу бизнес-приложений не только для коммерческих сфер, таких как розничная торговля, банковского дело, телекоммуникации, страхование, но и для специфичных сфер – медицина, молекулярная генетика и генная инженерия, Прикладная химия.

Архитектура вычислительной системы и модели программирования напрямую зависят от используемых алгоритмов и наборов данных.

Системы с общей памятью эффективно используются для поддержания большого числа обменов сообщениями между ветками параллельной программы, максимально уменьшая задержки между узлами, но накладывает ограничения на объем обрабатываемого массива данных. Для повышения их эффективности можно использовать коммутационные сети - гибридные сети с общей и распределенной память. Системы с распределенной памятью могут эффективно использовать на вычислительных узлах возможности оперативной памяти, но стоит сократить до минимума обмен сообщениями между узлами. В то же время объемы данных на узлах и отказоустойчивость так же нуждаются в контроле, так как при сбое, не выгруженные в постоянное хранилище данные будут утеряны.

Рассмотрим тенденции и направления развития в сфере Data Mining:

Augmented analytics (расширенная аналитика) развивается быстрыми темпами, объединяя анализ данных с алгоритмами машинного обучения и обработкой естественного языка, что дает возможность взаимодействовать с данными и эффективнее взаимодействовать с ними, выявляя ценные и необычные тенденции. Поскольку 80% работы ученых связано со сбором и переработкой информации, предполагается, что метод поможет ученым сосредоточится на более продуктивных задачах. Некоторые эксперты считают, что в 2020 году расширенная аналитика станет основным приобретением предприятий, занимающихся аналитикой и бизнес-аналитикой.

Облачные вычисления также были с нами в течение многих лет, и мы находим все больше и больше применений этой технологии. Но мы приблизились к границе возможности их оптимизации, что позволило найти возможный способ оптимизации cold storage (холодное хранение), а также стратегии гибридного облака (сочетание частного облака и сторонних, общедоступных облачных сервисов) и нескольких облаков (смесь различных инструментов и решений, доступных в разных облаках).

Continuous Intelligence (CI) – ожидаемая тенденция 2020 года, объединяющая аналитику в реальном времени с бизнес-операциями. Эта технология обрабатывает исторические и текущие данные одинаково, чтобы

поддержать процесс принятия решений или, в некоторых случаях, автоматизировать их.

CI использует другую технологию, упомянутую в этой статье – расширенную аналитику. Непрерывный интеллект – это новая технология, которая стала возможной благодаря расширенной аналитике и развитию других технологий больших данных и искусственного интеллекта. Потенциальные приложения еще впереди, но мы можем предсказать, что CI может помочь в:

- Обеспечение более эффективной поддержки клиентов.
- Разработка специальных предложений и скидок с учетом потребностей и ожиданий каждого клиента.
- Оптимизация корпоративного процесса принятия решений.

Gartner прогнозирует, что к концу 2022 года более 50 % новых бизнес-систем будут использовать непрерывный интеллект.

В тоже время в технологиях Data Mining существует множество проблем, которые имеют научное значение и еще не решены полностью. К таким проблемам относятся задачи подготовки исходных данных для каждой конкретной задачи, не важно в какой предметной области эта задача ставится. Например, в задачах скоринга – предсказательных моделях оценки качества. Такие задачи характерны для банковской сферы, области HR, медицины и т. д. [9].

Например, хорошо известная сфера банковского кредитования использует технологии Data Mining для ответа на вопрос, является ли заемщик кредитоспособным. Оценка надежности заемщика проводится по совокупности признаков, называемых категориальными переменных.

Подготовка набора категориальных переменных в качестве предсказателей представляет проблему для аналитика, особенно когда они демонстрируют высокую степень разброса (большое количество различных значений) [8].

Численные модели (например, линейная регрессия и большинство нейронных сетей) не могут принимать эти переменные непосредственно в качестве входных данных, поскольку операции между категориями и числами не определены. Определение – это набор решающих правил, которые также задаются на стадии предварительной подготовки модели. Иногда выгодно (даже необходимо) перекодировать такие переменные, как одну или несколько числовых фиктивных переменных, при этом каждая новая переменная содержит значение 0 или 1, указывающее на наличие (1) или отсутствие (0) одного отдельного значения. Это часто хорошо работает с небольшими и средними выборками. Однако по мере увеличения кардинальности фиктивные переменные быстро становятся неудобными: они добавляют потенциально много параметров к модели, уменьшая при этом среднее количество наблюдений, доступных для установки каждого из этих параметров.

Одним из решений является преобразование каждой отдельной категориальной переменной в новую, численную переменную, представляющую усредненное значение целевой переменной. В качестве примера можно привести замену переменной Доход на усреднение значения по выборке для определенной категории потенциальных заемщиков, таких как категория «домохозяйки».

Но, получение усредненных значений для категориальных переменных сталкивается с такой проблемой, как нехватка данных, пропущенные значения в выборке. Чтобы преодолеть это ограничение, можно выбрать либо n наиболее частых категорий, либо n категорий с самыми большими и маленькими значениями. Хотя оба этих метода иногда работают хорошо, они, в действительности, являются грубыми методами, особенно в задачах классификации, которые требуют обнаружения категорий с наиболее значительными отклонениями от глобального среднего.

Авторы предлагают метод, который заключается в том, чтобы выбрать только те категории, средние целевые значения которых, по крайней мере, минимальны кратным стандартным ошибкам отклонения от глобального среднего.

Подробная процедура заключается в следующем:

1. Рассчитать глобальное среднее (среднее целевой переменной, по всем наблюдениям)

2. Рассчитать локальное среднее для каждой категории.

a. Рассчитать локальное среднее и локальное стандартное отклонения средней для целевой переменной).

b. Рассчитать абсолютное значение разницы между локальным средним и глобальным средним значением и разделите ее на ошибку стандартного отклонения. Обычно она составляет 0.05.

с. Задать порог, относительно которого принимается решение о классификации. В задачах бинарной классификации в сфере банковского кредитования, например, это ответ на вопрос надежный заемщик или нет. Если при расчете заданный порог превышен, то ответ – «нет», если значение попадает в доверительных интервал, то ответ – «да».

3. Все переменные, не попавшие в результирующую выборку, собираются в "другой" категории, и все такие наблюдения характеризуются, как неопределенные значения, по которым нельзя получить достоверного ответа.

Достоинство такого подхода в том, что не надо проводить процедуру интерполяции для получения недостающих значений в выборке.

Недостаток такого подхода в том, что трудно задать ошибку отклонения от глобального среднего и получить доверительный интервал попадания в классификацию. Данный подход требует эмпирического подтверждения. Авторы планируют провести верификацию модели с помощью программных инструментальных средств.

Предварительная подготовка данных в технологиях Data Mining является важнейшей частью всей работы и занимает много времени.

Для аналитика данных – это творческая многогранная работа, от которой зависит успех моделирования. В тоже время, в литературе мало уделяется внимания этому вопросу, что, с точки зрения авторов, является упущением. Авторы планируют продолжить работу в этом направлении.

ЛИТЕРАТУРА

1. Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San. Francisco, 2000 [Электронный ресурс]. Режим доступа http://media.wiley.com/product_data/excerpt/24/04712285/0471228524-1.pdf (Дата обращения 13.05.20).

2. MIKE2.0, Big Data Definition [Электронный ресурс]. Режим доступа: http://mike2.openmethodology.org/wiki/Big_Data_Definition (Дата обращения 13.05.20).

3. Луньков А. Д. Интеллектуальный анализ данных [Электронный ресурс] / А. Д. Луньков, А. В. Харламов // Саратовский национальный исследовательский университет. – Режим доступа: http://elibrary.sgu.ru/uch_lit/1141.pdf (Дата обращения 15.05.20).

4. Keith D. Foote, Big Data Trends in 2020 [Электронный ресурс]. Режим доступа: https://www.dataversity.net/big-data-trends-in-2020 (Дата обращения 14.05.20).

5. Введение в современный Data Mining [Электронный ресурс]. Режим доступа: https://statistica.ru/local-portals/data-mining/ (Дата обращения 12.05.20).

7. GridMiner: An Infrastructure for Data Mining on Computational Grids Peter Brezany1 , Jürgen Hofer1 , A Min Tjoa2 , Alexander Wöhrer1 1 Institute for Software Science, University of Vienna Liechtensteinstrasse 22, A-1090 Vienna, Austria, [электронный ресурс]. Режим доступа http://www.gridminer.org/publications/brezany_apac03.pdf (Дата обращения 13.05.20).

8. Лавлинская О. Ю. Решение задачи классификации данных на основе многослойного персептрона / О. Ю. Лавлинская, В. О. Логвина // Вестник Воронежского института высоких технологий. – № 2 (29). – 2019. – С. 59-64.

9. Dorian Pyle. Data Preparation for Data Mining. – Los Altos, California: Morgan Kaufmann Publishers, 1999.

DATA PREPARATION FOR DATA MINING

2020 O. U. Lavlinskaya, V. I. Kiseleva

Voronezh Institute of High Technologies (Voronezh, Russia)

The article provides an analytical overview of such approaches as Data Mining, KDID. The trends and prospects of Data Mining are examined. Attention is paid to the issue of preparing source data in mining models, an author's approach to preparing source data is proposed, based on the calculation of global and local values of the average for a classification sample of data.

Keywords: Data Mining, Big Data, data processing, decision visualization, decision trees, Neural networks, KDID, data preparation.