Рис. 6. а) Исходная матрица изображения; б) Матрица длин серий яркости для направления $\varphi = 0$

Обозначим N_g – число возможных значений яркости, N_r – число возможных длин серий, а $p(i, j)$ – число серий длины j с яркостью i . Выделяют [2] следующие статистики величины $p(i, j)$ для анализа текстур: SRE — Short Run Emphasis (Обратные моменты возрастают при коротких сериях), LRE — Long Run Emphasis (Моменты увеличены при длинных сериях), GLNU — Gray Level Non-uniformity (Неоднородность яркости), RLNU — Run Length Non-uniformity (Неоднородность длины серии), RPC — Run Percentage (Доля изображения в сериях).

Таким образом, можно акцентировать различные свойства текстуры и использовать их в качестве ее признаков.

Реализация совокупности представленных в работе методов позволяют достичь 75% точности определения типа материала для следующего набора исследуемых текстур: дерево, ткань, вязаная ткань, кирпичная кладка, стекло, металл, золото, ржавчина, кожа, пластик. При этом наибольшая часть неверных ответов составляют ошибки в парах дерево-ткань и металл-стекло. Данную проблему можно решить, разработав алгоритм определения изотропности материала. Изотропные материалы имеют одинаковые свойства во всех направлениях, анизотропные – разные. Изотропными материалами являются металлы, камни, бетон, некоторые пластмассы. Текстура ткани со сниженной глубиной цвета также будет иметь высокое значение изотропности. Текстура же дерева в большинстве своем является анизотропной. Таким образом, ошибки в определении типа текстуры в парах дерево-ткань будут снижены. Предположительно, определение изотропности повысит качество оценки и в проблемной паре металл-стекло [3].

Существуют и другие методы оценки текстурных свойств, такие как метод бегущих длин, метод LBP-гистограмм, и прочие. Используя их, можно также повысить вероятность правильных ответов нейросети.

1. Скобцов Ю.А., Родин Ю.В., Оверко В.С. Моделирование и визуализация поведения потоков крови при патологических процессах. — Донецк: Издатель Заславский А.Ю., 2008. — 212 с.
2. И.Л. Ковалева. Текстурные признаки изображений. Минск: БНТУ. – 2010., 26 с.
3. Сорокин А. Б., Зикеева Е.А. Распознавание объектов с использованием искусственных нейронных сетей по текстуре материала // Научно-технический вестник Поволжья, 2018, №3. С. 143 – 145.

Тучкова А.С., Кондрашева П.П.

Термин «Data mining». Задачи, решаемые методами data mining

*Поволжский государственный университет телекоммуникаций и информатики
(Россия, Самара)*

doi: 10.18411/lj-10-2019-26

idsp: ljjournal-10-2019-26

Аннотация

Прогресс в технологии сбора и хранения цифровых данных привел к росту огромного количества баз данных. Это произошло во всех сферах человеческой

деятельности, от мирского (например, данные о транзакциях в супермаркете, записи об использовании кредитных карт, данные телефонных звонков, и государственная статистика) к более экзотическим (например, изображения астрономических тел, молекулярные базы данных и медицинские записи). Неудивительно, что этот интерес вырос в возможности использования этих данных, извлечения из них информации, которая может быть значима для владельца базы данных. Дисциплина, связанная с этой задачей, известна, как интеллектуальный анализ данных, о чем мы и говорим в статье.

Ключевые слова: Data Mining, интеллектуальный анализ данных, ПО, базы данных

Abstract

Advances in digital data collection and storage technology have led to the growth of a huge number of databases. This happened in all areas of human activity, from mundane (for example, supermarket transaction data, credit card usage records, phone call data, and government statistics) to more exotic ones (for example, images of astronomical bodies, molecular databases and medical records). Not surprisingly, this interest has grown in the possibility of using this data, extracting from it information that may be significant for the owner of the database. The discipline associated with this task is known as data mining, which we discuss in the article.

Keywords: Data Mining, database, software

Введение

Организации по всему миру генерируют большое количество данных, которые в основном мало сформированы или же несформированы. Эти нескомпанованные данные требуют обработки для получения значимой и полезной информации. Чтобы организовать большой объем данных, необходимо реализовать концепцию систем управления БД, таких как Oracle и SQL Server. Эти системы управления базами данных требуют использования SQL, специализированного языка запросов для извлечения информации из базы данных. Однако использование SQL не всегда адекватно для удовлетворения потребностей конечных пользователей в специализированной и сложной информации из несформированного большого банка данных. Это требует поиска определенных альтернативных методов для извлечения информации из большого и неорганизованного источника, поэтому и начинают использовать «Data Mining».

Задачи, решаемые методами интеллектуального анализа данных

Так что же такое «Data mining»? Это процесс извлечения знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т.д. Полученные знания должны быть достоверными, полезными и интерпретируемыми. Data Mining берет свое начало в различных дисциплинах, из которых двумя наиболее важными являются статистика и машинное обучение. Статистика имеет корни в математике — это раздел науки, который занимается сбором и анализом числовых данных с использованием различных методов и приемов. Использование данного раздела возможно на разных этапах интеллектуального анализа данных. Статистика может быть задействована в: стадии очистки данных, этапе сбора данных, анализа данных, а вот машинное обучение большинство своих истоков имеет из компьютерной области. В процессе интеллектуального анализа данных используются различные инструменты анализа, чтобы определить взаимосвязи между данными в хранилищах, и использовать их для правильного прогнозирования.

Интеллектуальный анализ данных - это интеграция различных методов из различных дисциплин, таких как статистика, машинное обучение, распознавание образов, нейронные сети, системы обработки изображений, управления базами данных и так далее, как и утверждалось ранее. Различные методы data mining характеризуются

определенными свойствами, которые могут сделать решающий выбор при поиске метода анализа данных. Методы можно сравнивать друг с другом, оценивая характеристики их свойств. Среди основных свойств и характеристик методов интеллектуального анализа данных рассмотрим такие, как точность, масштабируемость, интерпретируемость, проверяемость, сложность, гибкость, скорость и популярность. Большинство аналитических методов, используемых в технологии интеллектуального анализа данных, являются известными математическими алгоритмами и методами. Новое чертой в применении аналитических методов является возможность их использования при решении определенных конкретных задач, за счет возникающих возможностей аппаратного и программного обеспечения. Следует отметить, что большинство методов интеллектуального анализа данных были разработаны в рамках теории искусственного интеллекта.

Также укажем, что единого мнения относительно того, какие задачи следует относить к Data Mining, не наблюдалось, пока мы изучали материал для подготовки данной статьи, но большинство источников перечисляют следующие: классификация, кластеризация, прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов. Первое – это то, что возможна группировка фрагментов определенных данных вместе, что и есть классификация. Эти объединения должны иметь «естественные» отношения друг с другом для того, чтобы можно было провести более эффективный анализ. Следующее свойство – делать прогнозирование, основанное на математической теории вероятности. Их часто называют доверительными, поэтому они в основном измеряются. Прогнозирование интеллектуального анализа данных также может указывать на условия для того, чтобы поддержать теорию о потенциальных будущих объемах минимальной логической операции. А также визуализация данных — это предоставление информации с помощью рисунков, графиков, диаграмм, таблиц и диаграмм. Ценность визуализации заключается в том, что она часто позволяет наиболее четко и кратко идентифицировать и отображать информацию, содержащуюся в данных, поскольку изображение может сразу показать, что займет более одного абзаца в словесном эквиваленте.

В разработке методов визуального представления данных в равной степени участвуют ИТ-специалисты и дизайнеры, так как дизайн во многом зависит от того, насколько понятна и «читаема» визуализация.

С помощью визуализации данных решаются самые разные задачи. Во-первых, это важный инструмент на начальных этапах анализа данных. Самая простая графика позволяет быстро найти закономерности, тренды или аномалии данных, которыми будет руководствоваться аналитик при дальнейшей работе с данными. Аналогичным образом, журналист, используя графики при просмотре первичных данных, может сформулировать для себя вопросы, из которых в дальнейшем может привести сюжет для другого материала. Во-вторых, визуализация часто играет важную роль в представлении окончательных результатов анализа. Это могут быть статические графики, показывающие тренды; интерактивная визуализация, позволяющая пользователям исследовать данные; и инфографика (статическая и интерактивная), четкое изображение основано на истории данных. Важной задачей визуализации является то, что на одном изображении с ограниченным количеством физических измерений (обычно двухмерных) отображаются данные, содержащиеся во множестве измерений и взаимосвязи между ними. Методы интеллектуального анализа данных дают уверенность в отношении предсказанных решений с точки зрения согласованности прогноза и частоты правильных прогнозов.

Интеллектуальный анализ данных необходим для выявления существующих связей в наборе данных. Фактическая задача data mining – это автоматический или полуавтоматический анализ больших объемов данных для извлечения ранее неизвестных интересных шаблонов, таких как группы записей данных (кластерный

анализ), необычные записи (обнаружение аномалий) и зависимости (извлечение правил ассоциации). Это обычно включает использование методов базы данных, таких как пространственные индексы. Эти шаблоны затем можно рассматривать как своего рода сводку входных данных, и их можно использовать в дальнейшем анализе или, например, в машинном обучении и прогнозной аналитике. Например, этап интеллектуального анализа данных может идентифицировать несколько групп в данных, которые затем можно использовать для получения более точных результатов прогнозирования системой поддержки принятия решений. Ни сбор данных, ни подготовка данных, ни интерпретация результатов и создание отчетов не являются частью этапа анализа данных, связанные термины «драгирование данных», «промысел данных» и «отслеживание данных» относятся к использованию методов интеллектуального анализа данных для выборки частей большого набора данных о населении, которые (или могут быть) слишком малы, чтобы можно было сделать надежные статистические выводы о достоверности любых данных. шаблоны обнаружены. Эти методы, однако, могут быть использованы при создании новых гипотез для проверки больших массивов данных.

Заключение

И в конце хотелось бы отметить, что область использования интеллектуального анализа данных ничем не ограничена - он везде, где есть большое количество информации.

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. БХВ-Петербург, 2004
2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. Питер, 2013
3. Бергер А. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных. БХВ-Петербург, 2007
4. Макленнен Д., Танг Ч., Криват Б. Microsoft SQL Server 2008: Data Mining — интеллектуальный анализ данных . БХВ-Петербург, 2009
5. Чубукова И. А. Data Mining. Интернет-Университет Информационных Технологий (ИНТУИТ), 2016
6. Мастицкий С.Э., Шитиков В.К., Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Тольятти, 2017
7. Сегаран Т. Программируем коллективный разум . Символ-Плюс, 2008
8. Mitra S., Acharya T., Data Mining. Multimedia, Soft Computing, and Bioinformatics. – John Wiley & Sons, Inc. – Hoboken, New Jersey. –2003.
9. Konar A., Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain. – CRC Press LLC. – Boca Raton, Florida/ – 2000.
10. Han J., Kamber M., Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2001
11. [Electronic resource] URL: https://habr.com/ru/hub/data_mining/ (accessed: 15.09.2019)

Komlev G.V., Mitrofanova A.S.
Data mining models and methods

Reshetnev Siberian State University of Science and Technology
(Russia, Krasnoyarsk)

doi: 10.18411/lj-10-2019-27

idsp: ljjournal-10-2019-27

Аннотация

В этой статье описаны модели, методы интеллектуального анализа данных и возможные варианты их использования: базовая архитектура моделей и структур, свойства моделей интеллектуального анализа данных.

Ключевые слова: интеллектуальный анализ, модели, методы.

Abstract

This article describes the models, data mining methods and possible options for their use: the basic architecture of models and structures, the properties of data mining models.

Keywords: intellectual analysis, models, methods.