Инженерный вестник Дона, №12 (2021)
ivdon.ru/ru/magazine/archive/n12y2021/7338

Методы объединения, сокращения размеров и обработка больших данных

П.И. Питкевич

Белорусский государственный университет информатики и радиоэлектроники

Аннотация: Актуальность исследования обусловлена тем, что анализ больших данных может быть проблематичным, так как он зачастую включает сбор и хранение смешанных данных, которые основаны на разных правилах или закономерностях. В связи с этим, данная статья рассматривает анализ существующих методов обработки больших данных, которые можно применить к обработке смешанных или разнородных данных. В статье описываются преимущества и недостатки самых часто применяемых методов обработки смешанных данных. Раскрываются проблемы обработки разнородных данных. Приведены инструменты обработки больших данных, некоторые традиционные методы интеллектуального анализа данных, а также машинного обучения. Представлены преимущества слияния больших смешанных данных. В данной работе под разнородными данными следует понимать любые данные с высокой изменчивостью типов данных, форматов и характера происхождения. Материалы статьи представляют практическую ценность для обработки больших данных, выбора методов обработки больших данных, включая очистку данных, объединение данных, сокращение размеров и обработки смешанных данных и соответствующего аналитического и системного анализа.
Ключевые слова: разнородные данные, смешанные данные, разномасштабные данные, методы обработки данных, интеллектуальный анализ, аналитика данных.

Введение

Разнородные данные — это данные, обладающие высокой изменчивостью типов, которые отличаются характером происхождения и форматом [1]. Такие данные, по причине недостоверности или высокой избыточности, а также при наличии пропущенных значений могут быть неоднозначными и низкого качества. Поэтому встаёт серьёзный вопрос об агрегации и интеграции таких данных для последующей обработки, либо при принятии конкретных решений. К примеру, очень часто можно встретить ситуации, когда из Интернета вещей генерируются разнородные данные.

Для будущих запросов решающее значение имеют метаданные. Для некоторых документов на расширяемом языке разметки (XML) и реляционных таблиц, явные определения схемы в языке структурированных запросов (SQL), а также определение схемы XML или определение типа

документа можно получить напрямую из источников, и они интегрированы в метамодель. Для перевода данных применяется техника XML. Сложной частью являются полуструктурированные данные (например, JSON, XML или частично структурированные файлы CSV или Excel), содержащие неявные схемы.

Вопросы, связанные с управлением метаданными, чрезвычайно важны. Чтобы правильно интерпретировать разнородные данные, необходимо иметь подробные метаданные. Некоторые отчеты включают в себя только часть метаданных, однако для исследовательских целей необходимо значительно больше деталей, к примеру, о конкретном датчике, применяемом для сбора данных.

Создание алгоритмов обработки больших данных сфокусировано на решении проблем, появляющихся в связи с распределением таких данных, имеющих сложные и динамические характеристики.

Материалы и методы

Очисткой данных называется процесс определения неточных, неполных или необоснованных данных, с последующим изменением или удалением таких данных для того, чтобы улучшить качество данных [2]. Так как качество данных влияет на качество информации, которое в свою очередь влияет на процесс принятия решений, необходимо создать эффективные подходы к очистке больших данных, для улучшения качества данных с точки зрения принятия эффективных и точных решений.

Отсутствующее значение для переменной представляет собой значение, которое отсутствует в наборе данных. Пропущенные значения в переменной заменяются одним значением, к примеру, средним значением или медианой. Но это может вызвать снижение точности результатов обработки, потому что стандартные ошибки недооцениваются, значение корреляции между переменными искажается, и в статистических тестах

могут выдаваться неверные значения. Для большей части проблем с отсутствующими данными такого рода подхода нужно избегать. Можно попробовать исследовать корреляции между переменными с неизвестными и номинальными переменными. Неизвестные значения можно заполнить посредством исследования более точных корреляций. Всякий раз, когда требуется обработать набор данных с пропущенными значениями, можно использовать такие приемы, как: заполнение неизвестных значений, при помощи корреляции между переменными; заполнение неизвестных значений, путем сходства между данными (сопоставляются значения до и после); удаление данных с неизвестными [3].

Базы данных могут также иметь нерелевантные атрибуты. Значит, анализ релевантности в форме корреляционного анализа и выбора подмножества атрибутов можно применять для выявления атрибутов, которые не значимы для задачи прогнозирования или классификации. В противном случае, включение подобных атрибутов может затормозить и, вероятно, ввести в заблуждение этап обучения модели. Обычно, очистка данных и интеграция данных осуществляются как этап предварительной обработки. Несоответствие в именовании измерений или атрибутов может вызвать избыточность в результирующем наборе данных. Зачастую удаление избыточных данных рассматривается как основа очистки данных, в том числе сокращение данных.

В случае агрегирования или интеграции наборы данных сравниваются и соединяются на основе общих атрибутов и переменных. Усовершенствованные методы обработки и анализа данных дают возможность комбинировать как неструктурированные, так и структурированные данные для получения новых методов, подходов, идей; но для этого необходимы «чистые» данные. Методы объединения данных применяются для сравнения и агрегации, для создания или улучшения

представления реального положения дел, что помогает осуществить качественный анализ данных. Имеющиеся методологии объединения данных среднего уровня, объединяющие структурированные данные, в основном функционируют хорошо. Тем не менее, задачи по объединению данных высокого уровня для объединения нескольких неструктурированных данных с различных датчиков так и остаются весьма сложными.

Развитие инструментов интеграции данных ведётся в сторону унификации неструктурированных и структурированных данных. Зачастую необходимо структурировать неструктурированные данные и соединять в единый уровень данных разнородные источники и типы информации.

Большая часть платформ интеграции данных базируется на использовании первичной модели интеграции, которая основана на XML-типах или реляционных данных.

Для интеграции структурированных и неструктурированных данных можно использовать следующие подходы [4]:

– Применение для интеграции неструктурированных и открытых данных. В открытых наборах данных объекты можно использовать для идентификации именованных объектов (мест, организаций, людей), в свою очередь используемых для категоризации и организации текстового содержимого. Для задач связывания неструктурированных и структурированных данных можно применять инструменты распознавания и связывания именованных объектов, к примеру, DBpedia Spotlight.

– Распознавание и связывание сущностей. Фундаментальным шагом является извлечение структурированной информации из неструктурированных данных. Часть проблемы можно решить при помощи методов извлечения информации, например: извлечение онтологий, распознавание сущностей.

– Конвейеры обработки естественного языка. Можно использовать для проектов, которым необходима работа с неструктурированными данными.

При объединении данных из разнородных мультисенсорных систем есть 3 источника ошибок: несовместимости в определениях объектов, несовместимости типов данных, ошибки ввода данных. Очень часто многие предприятия для интеграции данных применяют извлечение, преобразование, загрузку, а также хранилища данных. Но в последнее время для интеграции данных технология, которая известна, как виртуализация данных, нашла некоторое признание в роли альтернативного решения. Виртуализация данных представляет собой объединенную базу данных, именуемую ещё составной базой данных. Виртуализация данных и стандартизация корпоративных данных приводят к снижению стоимости и времени внедрения интеграции данных. В отличие от хранилища, виртуализация данных выполняет очистку данных, объединение и преобразование данных, программно с применением логических представлений. Виртуализация данных допускает повторное использование и расширяемость, допуская цепочку логического представления. В основном стандартизация корпоративных данных даёт возможность уйти от проблемы несовместимости данных и несоответствия типов данных. Вместе с тем, виртуализация данных — это не замена хранилища данных; виртуализация данных позволяет уменьшить определенные аналитические нагрузки с хранилища данных.

Без хранилища данных не обойтись при регрессионном анализе, работе с многомерной структурой данных и анализе больших объемов данных [5].

Озера данных — это новый и мощный подход к решению задач интеграции данных, так как предприятия расширяют доступ к облачным и мобильным приложениям, а также Интернету вещей на основе разных

сенсоров. Это огромное хранилище, в котором различные данные хранятся в «сыром», т.е. в необработанном и неупорядоченном виде. Видеоролики, журналы, книги, фотографии и аудиозаписи, документы PDF и Word — это все неструктурированные данные, и все они могут храниться в озере данных. Не имеет никакого значения источник данных. Озеро данных может принимать данные из ERP- или CRM-систем, банковских программ [6], продуктовых каталогов, умных устройств или датчиков. Подобные хранилища подходят больше для менее структурированных данных, но и при работе с ними могут появиться сложности, например: усложнение и расширение управления метаданными по сравнению с необработанными данными, полученными из разнородных источников данных; работа со структурными метаданными из источников данных и аннотирование данных и метаданных дополнительной информацией, во избежание двусмысленности.

Так как структура подобных данных в озерах данных неизвестна, то без описания хранящихся данных, метаданных или моделей управления этими данными - их последующая обработка будет усложнена, потому что в таком виде все данные будут храниться хаотически, и представлять собой набор каких-то данных, без чёткого назначения и интереса [7].

Сокращение размеров и обработка больших данных

Существуют причины для уменьшения размерности данных. Так, для данных большого размера необходимы большие вычислительные мощности. Высокая размерность приводит к плохим способностям обобщения алгоритма обучения в некоторых ситуациях. Уменьшение размерности можно применить для поиска значимой структуры данных, интерпретируемости данных и целей иллюстрации [8].

Выбор подмножества функций — это известная задача машинного обучения и интеллектуального анализа данных. Генетические алгоритмы

очень часто применяются для решения задач выбора подмножества объектов. Уменьшение размерности, которое обеспечивается процессом подмножества функций, может дать несколько преимуществ: более быстрое введение окончательной модели классификации; повышение точности классификации; улучшение понятности окончательной модели классификации.

Методы выбора признаков делятся на 2 вида.

При первом подходе ранжирование объектов осуществляется по некоторым критериям, после чего подбираются объекты выше определенного порога. Второй подход делится на 3 части: подходы фильтра — сначала отбирают признаки, после чего применяют данное подмножество для осуществления алгоритма классификации; встроенные подходы - подбор признаков осуществляется как часть алгоритма классификации; и алгоритм определения по набору данных применяется для поиска оптимальных параметров.

В наборе данных с большим количеством переменных, обычно есть большое количество совпадений в информации. Простым способом отыскать данную избыточность является проверка корреляционной матрицы, полученной при помощи корреляционного анализа. После этого можно применять факторный анализ — метод снижения размерности, для того, чтобы понимать основные причины корреляции между группой переменных. Факторный анализ можно использовать для снижения количества переменных и определения структуры в отношениях между переменными. В связи с этим, факторный анализ часто применяется как метод установления структуры или сокращения данных [9]. Метод главных компонент полезен также, когда есть данные о большом количестве переменных и, вероятно, в этих переменных есть определённая избыточность. В таком случае избыточность подразумевает, что некоторые переменные коррелируют друг с

другом. Метод главных компонент очень эффективный, быстрый, простой и широко применяемый.

Некоторые моменты данного метода, которым следует уделить внимание:

Предварительная обработка. Исследование сложных моделей многомерных данных зачастую осуществляется очень медленно и в том числе подвержено переобучению. В модели число параметров, как правило, по количеству измерений экспоненциально. При помощи метода главных компонент выделяются элементы, которые перебалансируют вес данных, чтобы в некоторых случаях увеличить производительность вычислений [10].

Моделирование. Метод главных компонент иногда применяется как целая модель, к примеру, при предварительном распределении для новых данных [11].

Сжатие. Метод главных компонент можно применять для сжатия данных, заменяя данные на их низкоразмерное представление. Главные этапы применения факторного анализа или метода главных компонент состоят в следующем: подготовка данных; выбор фактор-модели, с определением того, что лучше подходит для целей исследования –метод главных компонент или факторный анализ, и подбор конкретного метода факторинга. Ели выбран подход факторного анализа, необходимо решить, сколько компонентов/факторов извлечь и произвести извлечение. Что касается выбора количества компонентов для извлечения, есть критерии для установления количества компонентов. Они включают в себя: основание количества компонентов на предыдущем опыте и теории; выбор количества компонентов, которые необходимы, чтобы учесть некоторое пороговое значение совокупной величины дисперсии в переменных; выбор количества сохраняемых компонентов посредством исследования собственных значений матрицы корреляции среди переменных [12].

Заключение

Для некоторых алгоритмов необходимо, чтобы данные были нормализованы (стандартизированы), прежде чем алгоритм может быть эффективно реализован. Нормализация (или стандартизация) подразумевает замену каждой исходной переменной стандартизированной версией переменной, которая имеет единичную дисперсию. Эффект данной нормализации (стандартизации) заключается в том, чтобы дать всем переменным с точки зрения изменчивости равную важность, при этом данные зачастую нормализуются перед осуществлением метода главных компонент.

Литература

1. Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Москва. Манн, Иванов и Фербер. 2013. С 34.

2. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Springer, 2003. pp. 107-122.

3. Marz N., Warren J. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications; 1st edition. 2015. pp. 76-78.

4. Багутдинов Р.А., Саргсян Н.А., Краснопахтыч М.А. Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных. Экономика. Информатика. Белгород. НИУ «БелГУ». Издательский дом «БелГУ». 2020. Том 47, №4. С. 792-802.

5. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014. pp. 87-89.

6. Zhang J, Yang X, Appelbaum D. Toward effective Big Data analysis in continuous auditing. Accounting Horizons. New Jersey. Rutgers. The state university of New Jersey. 2015. P. 115.

7. Питкевич П.И. Методика повышения эффективности управления облачными ресурсами накопления и обработки данных в банковской сфере. // Инженерный вестник Дона, 2021, №10. URL: ivdon.ru/ru/magazine/archive/n10y2021/7239.

8. Gadepally V., Kepner J. Big data dimensional analysis. 2014 IEEE High Performance Extreme Computing Conference (HPEC). 2014. Pp. 1-6.

9. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. New Jersey. Wiley. 2015. p 256.

10. Luengo J, García-Gil D., Ramírez-Gallego S., García S., Herrera F. Big Data Preprocessing. 2020. Pp. 147-160.

11. Witten I.H., Frank E., Hall M.A., Pal C. Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems). 4th Edition. 2016. P. 67.

12. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. Москва. Фазис. 2005. С. 159.

References

1. Shenberger V., Kuk`er K. Bol`shie danny`e. Revolyuciya, kotoraya izmenit to, kak my` zhivem, rabotaem i my`slim. [Big Data: A Revolution That Will Transform How We Live, Work, and Think]. Moskva. Mann, Ivanov i Ferber. 2013. P. 34.

2. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Springer, 2000. pp. 107-122.

3. Marz N., Warren J. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications; 1st edition. 2015. pp 76-78.

4. Bagutdinov R.A., Sargsyan N.A., Krasnoplaxty`ch M.A. Ekonomika. Informatika. Belgorod. NIU «BelGU». Izdatel'skij dom «BelGU», tom 47, № 4. Pp. 792-802.

5. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014. 2014. Pp. 87-89.

6. Zhang J, Yang X, Appelbaum D. Toward effective Big Data analysis in continuous auditing. Accounting Horizons. New Jersey. Rutgers. The state university of New Jersey. 2015. P. 115.

7. Pitkevich P.I. Inzhenernyj vestnik Dona, 2021, №10. URL: ivdon.ru/ru/magazine/archive/n10y2021/7239.

8. Gadepally V., Kepner J. Big data dimensional analysis. 2014. IEEE High Performance Extreme Computing Conference (HPEC). 2014. pp 1-6.

9. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. New Jersey. Wiley. 2015. P. 256.

10. Luengo J, García-Gil D., Ramírez-Gallego S., García S., Herrera F. Big Data Preprocessing. 2020. pp. 147-160.

11. Witten I.H., Frank E., Hall M.A., Pal C. Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems) 4th Edition. 2016. p. 67.

12. Zhuravlev Yu.I., Ryazanov V.V., Sen`ko O.V. Raspoznavanie. Matematicheskie metody`. Programmnaya sistema. Prakticheskie primeneniya. [Recognition. Mathematical methods. Software system. Practical applications]. Moskva. Fazis. 2005. p. 159.