18+

Вадим Николаевич
Шмаль
Сергей Сергеевич
Павлов

Интеллектуальный анализ данных

Учебник

Вадим Николаевич Шмаль
Сергей Сергеевич Павлов

Интеллектуальный анализ данных. Учебник

http://www.litres.ru/pages/biblio_book/?art=68765871
ISBN 9785005944801

Аннотация

Sergey Pavlov, master Plekhanov Russian University of Economics. Vadim Shmal, Ph. D., associate professor Russian University of Transport (MIIT).

Содержание

Интеллектуальный анализ данных Учебник

Вадим Николаевич Шмаль
Сергей Сергеевич Павлов

ISBN 978-5-0059-4480-1
Создано в интеллектуальной издательской системе Ridero

Интеллектуальный анализ данных

Интеллектуальный анализ данных – это процесс извлечения и обнаружения закономерностей в больших наборах данных с использованием методов на стыке машинного обучения, статистики и систем баз данных, особенно баз данных, содержащих большие числовые значения. Это включает в себя поиск в больших объемах информации статистически значимых закономерностей с применением сложных математических алгоритмов. Собранные переменные включают значение входных данных, уровень достоверности и частоту гипотезы, а также вероятность обнаружения случайной выборки. Он также включает в себя оптимизацию параметров для получения наилучшего шаблона или результата, корректировку входных данных на основе некоторых фактов для улучшения конечного результата. Эти параметры включают в себя параметры для статистических средних, таких как размеры выборки, а также статистические показатели, такие как частота ошибок и статистическая значимость.

Идеальный сценарий для интеллектуального анализа данных состоит в том, что параметры находятся в порядке, что обеспечивает наилучшие статистические результаты с наиболее вероятными значениями успеха. В этом идеальном сценарии интеллектуальный анализ данных происходит в рамках закрытой математической системы, которая

собирает все входные данные для системы и выдает наиболее вероятный результат. На самом деле идеальный сценарий редко встречается в реальных системах. Например, в реальной жизни этого не происходит при получении инженерно-сметной документации по реальному дизайн-проекту. Вместо этого для расчета наилучшей оценки успеха используется множество факторов, таких как параметры проекта и текущая сложность приведения проекта в соответствие со спецификациями проекта, и эти параметры постоянно меняются по мере продвижения проекта. Хотя они могут быть полезны в определенных ситуациях, например при разработке конкретных продуктов, их значения должны подвергаться постоянной переоценке в зависимости от текущих условий проекта. На самом деле лучший анализ данных происходит в сложной математической структуре задач с множеством переменных и множеством ограничений, а не в закрытой математической системе всего с несколькими переменными и закрытой математической структурой.

Данные часто собираются из множества разных источников и нескольких разных направлений. Каждый тип данных анализируется, и все эти выходные данные анализируются, чтобы получить оценку того, как каждая часть данных может или не может быть вовлечена в конечный результат. Такой анализ часто называют процессом анализа или анализом данных. Анализ данных также включает в себя определение другой важной информации о базе данных, которая может

иметь или не иметь прямого влияния на результаты. Часто они также генерируются из разных источников.

Данные обычно собираются из множества различных источников, и для получения наилучших статистических результатов применяется множество статистических методов. Результаты этих методов часто называют статистическими свойствами или параметрами и часто задают математические формулы, которые предназначены для результатов каждой математической модели. Математические формулы часто являются наиболее важными аспектами процесса анализа данных и обычно структурируются с использованием математических формул, известных как алгоритмы. Некоторые математические алгоритмы основаны на некотором теоретическом подходе или модели. Другие математические алгоритмы используют логику и логические доказательства в качестве математических инструментов для понимания данных. Другие математические алгоритмы часто используют вычислительные процедуры, такие как математическое моделирование и математические инструменты, чтобы понять конкретную проблему или данные. Хотя такие вычислительные процедуры могут быть необходимы для завершения математической модели данных, такие математические алгоритмы могут иметь другие математические инструменты, которые могут быть более подходящими для реального мира. Хотя эти математические модели часто очень сложны, зачастую проще разработать математический алгоритм и модель

на основе математической модели, чем разработать математический алгоритм и модель на основе реального процесса анализа данных.

В действительности обычно существует ряд математических моделей, которые обеспечивают более полное понимание ситуации и данных, чем какая-либо одна математическая модель или математический алгоритм. Затем данные анализируются, и математическая модель данных часто используется для получения определенного значения параметра. Это значение параметра обычно определяется численными расчетами. Если параметр не имеет прямой связи с результатом окончательного анализа, параметр иногда рассчитывается косвенно с использованием статистической процедуры, которая дает параметр, имеющий прямую корреляцию с результатом анализа данных. Если параметр имеет прямую корреляцию с результатом анализа данных, этот параметр часто используется непосредственно для получения окончательного результата анализа. Если параметр не имеет прямого отношения к результату анализа, параметр часто получается косвенно с помощью математического алгоритма или модели. Например, если анализ данных может быть описан математической моделью, то параметр может быть получен косвенно с помощью математического алгоритма или модели. Обычно проще получить параметр прямо или косвенно с помощью математического алгоритма или модели.

Собирая и анализируя множество различных видов дан-

ных, а также выполняя математический анализ данных, данные можно анализировать, а статистику и другие статистические инструменты можно использовать для получения результатов. Во многих случаях использование численных расчетов для получения реальных данных может быть очень эффективным. Тем не менее, этот процесс обычно требует тестирования в реальных условиях перед анализом данных.

Агентный анализ данных

Агентный интеллектуальный анализ – это междисциплинарная область, которая объединяет многоагентные системы с интеллектуальным анализом данных и машинным обучением для решения бизнес-задач и решения проблем в науке.

Агенты можно описать как децентрализованные вычислительные системы, обладающие как вычислительными, так и коммуникационными возможностями. Агенты моделируются на основе алгоритмов обработки данных и сбора информации, таких как «проблема агента», которая представляет собой метод машинного обучения, который пытается найти решения бизнес-проблем без какого-либо центра обработки данных.

Агенты похожи на распределенные компьютеры, где пользователи совместно используют вычислительные ресурсы друг с другом. Это позволяет агентам обмениваться полезными данными и обрабатывать данные параллельно, эффективно ускоряя обработку и позволяя агентам быстрее выполнять свои задачи.

Обычным применением агентов является обработка и передача данных, например, задача поиска и анализа больших объемов данных из нескольких источников для определенных шаблонов. Агенты особенно эффективны, потому что у них нет централизованного сервера, который бы отслежи-

вал их действия.

В настоящее время в этой области существуют две технологии, которые обеспечивают ту же функциональность, что и агенты, но только одна из них широко используется: распределенные вычисления, которые основаны на ЦП и часто используют централизованные серверы для хранения информации; и локальные вычисления, которые обычно основаны на локальных устройствах, таких как ноутбук или мобильный телефон, при этом пользователи обмениваются информацией друг с другом.

Обнаружение аномалий

При анализе данных обнаружение аномалий (также обнаружение выбросов) – это идентификация редких элементов, событий или наблюдений, которые вызывают подозрения, поскольку значительно отличаются от большинства данных. Одним из применений обнаружения аномалий является безопасность или бизнес-аналитика как способ определения уникальных условий нормального или наблюдаемого распределения. Аномальные распределения отличаются от среднего тремя способами. Во-первых, они могут быть соотнесены с предыдущими значениями; во-вторых, существует постоянная скорость изменения (в противном случае они являются выбросом); и в-третьих, они имеют нулевое среднее значение. Регулярное распределение является нормальным распределением. Аномалии в данных могут быть обнаружены путем измерения среднего значения и деления на значение среднего значения. Поскольку не существует теоретического верхнего предела количества вхождений в наборе данных, эти множественные значения подсчитываются и представляют элементы, которые имеют отклонения от среднего, хотя они не обязательно представляют собой истинную аномалию.

Сходства аномалий данных

Понятие аномалии можно описать как значение данных,

которое значительно отличается от среднего распределения. Но описание аномалий также достаточно общее. В наборе данных может возникнуть любое количество отклонений, если существует разница между наблюдаемыми отношениями или пропорциями. Эта концепция наиболее известна для наблюдения за отношениями. Они усредняются для получения распределения. Сходство наблюдаемого соотношения или пропорции гораздо меньше аномалии. Аномалии не обязательно редки. Даже когда наблюдения более похожи, чем ожидаемые значения, наблюдаемое распределение не является типичным или ожидаемым распределением (выбросами). Однако существует также естественное распределение возможных значений, в которое могут вписаться наблюдения. Аномалии легко обнаружить, наблюдая за статистическим распределением наблюдаемых данных.

Во втором сценарии известное распределение отсутствует, поэтому невозможно сделать вывод, что наблюдения типичны для какого-либо распределения. Однако может быть доступное распределение, которое предсказывает распределение наблюдений в этом случае.

В третьем сценарии имеется достаточно различных точек данных, чтобы использовать полученное распределение для прогнозирования наблюдаемых данных. Это возможно при использовании данных, которые не являются очень нормальными или имеют разную степень отклонения от наблюдаемого распределения. В этом случае имеется среднее или

ожидаемое значение. Прогноз – это распределение, которое будет описывать данные, которые не являются типичными для данных, хотя они не обязательно являются аномалиями. Это особенно характерно для нерегулярных наборов данных (также известных как выбросы).

Аномалии не ограничиваются естественными наблюдениями. Фактически, большинство данных в деловой, социальной, математической или научной областях иногда имеют необычные значения или распределения. Чтобы помочь в принятии решений в таких ситуациях, можно выявить закономерности, относящиеся к различным значениям данных, отношениям, пропорциям или отличиям от нормального распределения. Эти закономерности или аномалии представляют собой отклонения, имеющие некоторое теоретическое значение. Однако значение отклонения обычно настолько мало, что большинство людей его не замечают. Его можно назвать аномальным значением, аномалией или разницей, причем любой из этих терминов относится как к наблюдаемым данным, так и к возможному основному распределению вероятностей, которое генерирует данные.

Проблемы оценки аномалий данных

Теперь, когда мы немного знаем об аномалиях данных, давайте рассмотрим, как интерпретировать данные и оценить возможность аномалии. Полезно рассматривать аномалии, исходя из предположения, что данные генерируются относительно простыми и предсказуемыми процессами. Следо-

вательно, если бы данные были сгенерированы конкретным процессом с известным распределением вероятностей, то мы могли бы уверенно идентифицировать аномалию и наблюдать за отклонением данных.

Маловероятно, что все аномалии связаны с распределением вероятностей, поскольку маловероятно, что некоторые аномалии связаны. Однако если есть какие-либо аномалии, связанные с распределением вероятностей, то это будет свидетельствовать о том, что данные действительно генерируются процессами или процессами, которые, вероятно, предсказуемы.

В этих обстоятельствах аномалия свидетельствует о вероятности обработки данных. Маловероятно, что закономерность отклонений или аномальных значений данных является случайным отклонением лежащего в основе распределения вероятностей. Это говорит о том, что отклонение связано с конкретным, случайным процессом. В соответствии с этим предположением аномалии можно рассматривать как аномалии данных, генерируемых процессом. Однако аномалия не обязательно связана с процессом обработки данных.

Понимание аномалии данных

В контексте оценки аномалий данных важно понимать распределение вероятности и ее вероятность. Также важно знать, распределена ли вероятность приблизительно или нет. Если она приблизительно распределена, то вероятность, скорее всего, будет примерно равна истинной вероятности.

Если оно не распределено приблизительно, то есть вероятность, что вероятность отклонения может быть немного больше, чем истинная вероятность. Это позволяет интерпретировать аномалии с возможностью большего отклонения как аномалии большей величины. Вероятность аномалии данных можно оценить с помощью любой меры вероятности, такой как вероятность выборки, правдоподобие или доверительные интервалы. Даже если аномалия не связана с конкретным процессом, все же можно оценить вероятность отклонения.

Эти вероятности необходимо сравнить с естественным распределением. Если вероятность намного больше естественной вероятности, то существует вероятность того, что отклонение не такой же величины. Однако маловероятно, чтобы отклонение намного превышало естественную вероятность, поскольку вероятность очень мала. Следовательно, это не свидетельствует о фактическом отклонении от распределения вероятностей.

Выявление значимости аномалий данных

В контексте оценки аномалий данных полезно определить соответствующие обстоятельства. Например, если есть аномалия в количестве задержанных рейсов, может случиться так, что отклонение будет довольно небольшим. Если задерживается много рейсов, более вероятно, что количество задержек очень близко к естественной вероятности. Если есть несколько рейсов, которые задерживаются, маловероятно,

что отклонение намного превышает естественную вероятность. Следовательно, это не будет свидетельствовать о значительно более высоком отклонении. Это говорит о том, что аномалия данных не имеет большого значения.

Если процентное отклонение от нормального распределения значительно выше, то есть вероятность, что аномалии данных связаны с процессом, как в случае с этой аномалией. Это является дополнительным свидетельством того, что аномалия данных является отклонением от нормального распределения.

После анализа значимости аномалии важно узнать, в чем причина аномалии. Связано ли это с процессом, сгенерировавшим данные, или не связано? Возникла ли аномалия данных в ответ на внешнее воздействие или она возникла внутри? Эта информация полезна при определении того, каковы перспективы получения дополнительной информации о процессе.

Причина в том, что не все отклонения связаны с изменчивостью процесса и по-разному влияют на процесс. В отсутствие понятного процесса определение влияния аномалии данных может оказаться сложной задачей.

Анализ важности аномалий данных

При отсутствии признаков отклонения от распределения вероятностей аномалии данных часто игнорируются. Это дает возможность выявить аномалии данных, которые имеют большое значение. В такой ситуации полезно рассчитать ве-

роятность отклонения. Если вероятность достаточно мала, то аномалией можно пренебречь. Если вероятность намного выше, чем естественная вероятность, то она может предоставить достаточную информацию, чтобы сделать вывод о том, что процесс имеет большую величину, а потенциальное воздействие аномалии имеет большое значение. Наиболее разумным предположением является то, что аномалии данных возникают часто.

Вывод

В контексте оценки точности данных важно выявить и проанализировать количество аномалий данных. Когда количество аномалий данных относительно невелико, маловероятно, что отклонение имеет значительную величину и влияние аномалии невелико. В этой ситуации аномалии данных можно игнорировать, но, когда количество аномалий данных велико, вполне вероятно, что аномалии данных связаны с процессом, который можно понять и оценить. В этом случае проблема заключается в том, как оценить влияние аномалии данных на процесс. Качество данных, частота данных и скорость, с которой генерируются данные, являются факторами, определяющими, как оценивать влияние аномалии.

Анализ аномалий данных имеет решающее значение для изучения процессов и повышения их производительности. Он предоставляет информацию о характере процесса. Эта информация может быть использована при оценке влияния отклонения, оценке рисков и преимуществ применения кор-

ректировок процесса. В конце концов, аномалии данных важны, потому что они дают представление о процессах.

Непрерывный процесс оценки воздействия аномалий данных предоставляет ценную информацию. Эта информация предоставляет полезную информацию о процессе и предоставляет лицам, принимающим решения, информацию, которую можно использовать для повышения эффективности процесса.

Этот подход дает возможность создавать аномалии данных, которые дают возможность оценить влияние аномалии. Цель состоит в том, чтобы получить представление о процессах и улучшить их производительность. В таком сценарии подход дает четкое представление о типе изменения процесса, которое может быть произведено, и о влиянии отклонения. Это может быть полезная информация, которую можно использовать для выявления аномалий процесса, которые можно оценить для оценки влияния отклонения. Процесс выявления аномалий процесса очень важен для получения ценных данных для оценки потенциальных аномалий в производительности процесса.

Анализ аномалий – это процесс, который оценивает частоту отклонений данных и сравнивает ее с фоновой частотой. Критерием оценки частоты отклонения данных является большее количество отклонений данных, а не естественное возникновение аномалий данных. В этом случае частота измеряется путем сравнения количества отклонений данных

с фоном возникновения отклонений данных.

Это предоставляет информацию о том, сколько отклонений данных вызвано процессом с течением времени и частотой отклонения. Это также может обеспечить связь с основным процессом отклонения. Эта информация может быть использована для понимания основной причины отклонения. Более высокая частота отклонения данных дает ценную информацию о процессе отклонения. В такой ситуации, вероятно, будет обнаружен риск отклонения и могут быть оценены необходимые изменения процесса.

Многие исследования проводятся по анализу аномалий данных для выявления факторов, способствующих возникновению аномалий данных. Некоторые из этих факторов относятся к процессам, которые требуют частых изменений процессов. Некоторые из этих факторов можно использовать для выявления процессов, которые могут быть аномальными. Многие параметры можно найти в системах, обеспечивающих характеристики процесса.

Изучение правила ассоциации

Изучение ассоциативных правил – это основанный на правилах метод машинного обучения для обнаружения интересных отношений между переменными в больших базах данных примеров. Эта техника вдохновлена слуховой системой, где мы изучаем правила ассоциации слухового стимула и только этого стимула.

Иногда при работе с набором данных мы не уверены, релевантны ли строки набора данных для задачи обучения, и если да, то какие. Мы можем захотеть пропустить те строки набора данных, которые не имеют значения. Следовательно, ассоциации обычно определяются неинтуитивными критериями, такими как порядок, в котором эти переменные появляются в последовательности примеров, или повторяющиеся значения в этих строках данных.

Этот проблематичный аспект изучения ассоциативных правил может быть устранен в виде алгоритма обнаружения аномалий. Эти алгоритмы пытаются обнаружить нестандартные шаблоны в больших наборах данных, которые могут представлять необычные связи между особенностями данных. Эти аномалии часто обнаруживаются алгоритмами распознавания образов, которые также являются частью алгоритмов статистического вывода. Например, изучение правил наивного Байеса может обнаруживать аномалии при изуче-

нии правил ассоциации на основе визуального осмотра представленных примеров.

В большом наборе данных пространство признаков может представлять область изображения как набор чисел, в котором каждый пиксель изображения имеет определенное количество пикселей. Характеристики изображения могут быть представлены в виде вектора, и мы можем поместить этот вектор в пространство признаков. Если пространство признака не пусто, признак будет числом пикселей в изображении, которые принадлежат определенному цвету.

Кластеризация

Кластеризация – это задача обнаружения групп и структур в данных, которые в той или иной мере «похожи», не используя известные структуры в данных, а обучаясь на том, что уже есть.

В частности, кластеризация используется таким образом, что новые точки данных добавляются только к существующим кластерам, без изменения их формы для соответствия новым данным. Другими словами, кластеры формируются до сбора данных, а не закрепляются после того, как все данные собраны.

Учитывая набор параметров для данных, которые (в основном) являются переменными, и их «коллинеарность», кластеризацию можно рассматривать как иерархический алгоритм для поиска кластеров точек данных, удовлетворяющих набору критериев. Параметры можно сгруппировать в одну из двух категорий: значения параметров, определяющие пространственное расположение кластеров, и значения параметров, определяющие отношения между кластерами.

Учитывая набор параметров для набора данных, кластеризацию можно рассматривать как обнаружение этих кластеров. Какие параметры мы используем для этого? Метод неявной кластеризации, который находит ближайшие кластеры (или, в некоторых версиях, кластеры, более похожие друг

на друга) с наименьшими вычислительными затратами, вероятно, является самым простым и наиболее часто используемым методом для этого. При кластеризации мы стремимся к тому, чтобы кластеры были как можно более связаны друг с другом – не имеет значения, делаем ли мы это, проводя больше измерений или используя только определенную технику для сбора данных.

Но в чем разница между кластеризацией и разделением данных на один или несколько наборов данных?

Методы неявной кластеризации и управляемой кластеризации на самом деле очень похожи. Вся разница в том, что мы используем разные параметры, чтобы определить, в каком направлении нам следует разделять данные. Возьмем в качестве примера набор точек на сфере, которые определяют взаимосвязанную сеть. Оба метода направлены на то, чтобы сеть была максимально близка к сети, определяемой двумя ближайшими точками. Это потому, что нам все равно, если мы очень далеко от одного или другого. Итак, используя алгоритм неявной кластеризации (кластерное расстояние), мы разделим сферу на две части, которые определяют очень разные сети: одна будет сетью, определяемой двумя ближайшими точками, а другая будет сетью, определяемой двумя самыми дальними точками. В результате получится две совершенно отдельные сети. Но это нехороший подход, потому что чем дальше мы удаляемся от двух ближайших точек, тем меньше расстояния между точками, тем труднее будет найти

связи между ними – так как существует ограниченное количество точек, которые связаны небольшим расстоянием.

С другой стороны, метод контролируемой кластеризации (кластерное расстояние) потребовал бы от нас измерения длины между каждой парой точек, а затем выполнения вычислений, которые делают ближайшие друг к другу сети наименьшим возможным расстоянием. Результатом, вероятно, будут две отдельные сети, которые близки друг к другу, но не совсем одинаковы. Поскольку нам нужно, чтобы две сети были похожи друг на друга, чтобы обнаружить взаимосвязь, вполне вероятно, что этот метод не сработает – вместо этого два кластера будут совершенно разными.

Различие между этими двумя методами сводится к тому, как мы определяем «кластер». Дело в том, что в первом методе (кластерное расстояние) мы определяем кластер как множество точек, принадлежащих сети, аналогичной сети, определяемой двумя ближайшими точками. По этому определению сети всегда будут связаны (они будут находиться на одинаковом расстоянии друг от друга), независимо от того, сколько точек мы включаем в определение. Но во втором методе (управление кластеризацией) мы определяем кластеры как пары точек, которые находятся на одинаковом расстоянии от всех других точек в сети. Это определение может сильно затруднить поиск связанных точек, потому что оно требует, чтобы мы находили каждую точку, аналогичную другим точкам в сети. Тем не менее, это понятный компро-

мисс. Сосредоточившись на поиске кластеров с одинаковым расстоянием друг от друга, мы, вероятно, получим больше полезных данных, поскольку, если мы найдем связи между ними, мы сможем использовать эту информацию, чтобы найти взаимосвязь между ними. Это означает, что у нас больше возможностей найти связи, что облегчит выявление отношений. Определяя кластеры с помощью измерений расстояния, мы гарантируем, что сможем найти взаимосвязь между двумя точками, даже если нет возможности напрямую измерить расстояние между ними. Но это часто приводит к очень малому количеству соединений в данных.

Глядя на пример создания двух наборов данных – один для неявной кластеризации и один для управляемой кластеризации – мы можем легко увидеть разницу между этими двумя методами. В первом примере результаты могут быть одинаковыми в одном случае и разными в другом. Но если метод хорош для поиска интересных взаимосвязей (как это обычно и бывает), он даст нам полезную информацию об общей структуре данных. Однако, если техника плохо выявляет взаимосвязи, то она даст нам очень мало информации.

Допустим, мы разрабатываем систему для определения направления нового продукта и хотим определить похожие продукты. Поскольку невозможно измерить направление продукта вне системы, нам придется найти связи между продуктами на основе информации об их названиях. Если есть хорошее правило, которое мы можем использовать для

установления связей между похожими продуктами, тогда эта информация очень полезна, поскольку она позволяет нам находить интересные отношения (путем идентификации похожих продуктов, которые появляются близко друг к другу). Однако, если связь между двумя продуктами не очень очевидна, вполне вероятно, что это просто несвязанная связь – а значит, выбранный нами метод обнаружения признаков может не иметь большого значения. С другой стороны, если связь не очень очевидна, но чрезвычайно полезна (как в приведенном выше примере), то мы можем начать узнавать, как название продукта связано с процессом, через который продукт прошел. Это пример того, как разные методы могут давать очень разные результаты.

В отличие от характеристик разных методов, у вас также есть разные возможные техники. Например, когда я говорю, что моя система использует распознавание изображений, это не обязательно означает, что процесс, через который проходит продукт, использует распознавание изображений. Если есть изображения продукта, которые мы сделали в прошлом, или если мы захватили некоторые входные данные из изображения продукта, полученная система, вероятно, не будет использовать распознавание изображений. Это может быть что-то совершенно другое – что-то гораздо более сложное. Каждый из этих методов способен идентифицировать очень разные вещи. Результат может зависеть от характеристик фактических данных или от используемых данных. Это озна-

чает, что недостаточно посмотреть на конкретный тип инструмента – нам также нужно посмотреть, какой тип инструмента будет использоваться для определенного типа процесса. Это пример того, как анализ данных не должен быть сосредоточен только на решаемой проблеме. Скорее всего, система проходит множество различных процессов, поэтому нам нужно посмотреть, как будут использоваться различные инструменты для создания взаимосвязи между двумя точками, а затем решить, какой тип данных рассматривать.

Часто мы будем больше озабочены тем, как будет применяться метод. Например, мы можем захотеть увидеть, какой тип данных, скорее всего, будет полезен для поиска связи. Мы видим, что нет большой разницы в том, как применяется обработка естественного языка. Это означает, что, если мы хотим найти взаимосвязь, обработка естественного языка будет хорошим выбором. Однако обработка естественного языка не решает все возможные отношения. Обработка естественного языка часто полезна, когда мы хотим сделать огромное количество маленьких шагов, но обработка естественного языка ничего не делает, когда мы хотим пойти действительно глубоко. Взгляд на обработку естественного языка позволяет устанавливать связи между данными, чего нельзя сделать при использовании других методов. Это одна из причин, по которой обработка естественного языка может быть полезной, но не необходимой.

Тем не менее, обработка естественного языка часто не на-

ходит таких сильных связей, как распознавание изображений, потому что обработка естественного языка фокусируется на более простых данных, тогда как распознавание изображений рассматривает очень сложные данные. В этом случае обработка естественного языка не очень хороша, но все же может быть полезна. Рассмотрение обработки естественного языка не всегда является лучшим способом решения проблемы. Обработка естественного языка может быть полезна, если данные простые, но иногда невозможно работать с очень сложными данными.

Этот пример можно применить ко многим различным типам данных, но обработка естественного языка, как правило, более полезна для данных естественного языка, таких как текстовые файлы. Для более сложных данных (таких как изображения) обработки естественного языка часто бывает недостаточно. Если есть проблема с обработкой естественного языка, важно рассмотреть другие методы, такие как определение слов и определение того, какие данные на самом деле хранятся в изображении. Этот тип данных потребует другой структуры данных, чтобы найти взаимосвязь.

С возрастающей сложностью технологий у нас часто нет времени просматривать данные, которые мы просматриваем. Даже если мы посмотрим на данные, мы можем не найти хорошего решения, потому что у нас есть большое количество вариантов, но не так много времени, чтобы рассмотреть их все. Вот почему во многих компаниях есть специа-

лист по данным, который может принять множество различных решений, а затем решить, что лучше всего подходит для данных

Классификация

Классификация – это задача обобщения известной структуры для применения к новым данным. Например, программа электронной почты может попытаться классифицировать электронное письмо как «законное», или как «спам», или, может быть, как «удаленное администратором», и если она сделает это правильно, то может пометить электронное письмо как актуально для пользователя.

Однако для серверов классификация более сложна, потому что хранение и передача находятся далеко от пользователей. Когда серверы потребляют огромные объемы данных, проблема в другом. Задача сервера состоит в том, чтобы создать хранилище и передать это хранилище, чтобы серверы могли получить к нему доступ. Таким образом, серверы часто могут избежать разглашения особо конфиденциальных данных, если они могут понять смысл данных при их поступлении, в отличие от обширных пулов данных, часто используемых для электронной почты. Проблема классификации отличается, и к ней нужно подходить по-другому, а существующие системы классификации для серверов не предоставляют интуитивно понятного механизма, позволяющего пользователям обрести уверенность в том, что серверы правильно классифицируют их данные.

Этот простой алгоритм полезен для классификации дан-

ных в базах данных, содержащих миллионы или миллиарды записей. Алгоритм работает хорошо, при условии, что все отношения в данных достаточно отличаются друг от друга и что данные относительно малы как в столбцах, так и в строках. Это делает классификацию данных полезной в системах с относительно небольшим объемом памяти и небольшим объемом вычислений, и поэтому классификация больших наборов данных остается серьезной нерешенной проблемой.

Простейшим алгоритмом классификации для классификации данных является метод полной корреляции, также известный как метод корреляции. При полной корреляции у вас есть два набора данных, и вы сравниваете данные одного набора с данными другого набора. Это легко сделать для отдельных фрагментов данных. Следующим шагом является вычисление корреляции между двумя наборами данных. Эта корреляция двух наборов данных говорит вам, какой процент данных составляет каждый набор. Таким образом, используя эту корреляцию, вы можете классифицировать данные либо как один набор, либо как другой, указывая на части набора данных, которые происходят из того или иного набора.

Этот простой метод часто хорошо работает для данных, хранящихся в простых базах данных с небольшим объемом данных и низкой скоростью доступа к данным. Например, система базы данных может использовать древовидную

структуру для хранения данных, при этом столбцы записи представляют поля в структуре. Эта структура не позволяла ранжировать данные, потому что данные находились бы в двух отдельных строках древовидной структуры. Это делает невозможным осмысление данных, если данные помещаются только в одну древовидную структуру. Если в базе данных есть два дерева данных, вам нужно будет сравнить каждое из двух деревьев. Если бы было большое количество деревьев, сравнение могло бы быть вычислительно затратным.

Следовательно, полная корреляция является плохим методом классификации. Корреляция данных не различает соответствующие части данных, и данные относительно малы как в столбцах, так и в строках. Эти проблемы делают полную корреляцию непригодной для простых систем классификации данных и систем хранения данных. Однако, если данные относительно велики, может применяться полная корреляция. Этот пример полезен для систем хранения данных с относительно высокой вычислительной нагрузкой.

Сочетание метода классификации данных с системой хранения данных повышает как производительность, так и удобство использования. В частности, размер результирующего алгоритма классификации в значительной степени не зависит от размера хранилища данных. Алгоритм подробной классификации вообще не требует много памяти для хранения данных. Часто он достаточно мал, чтобы хранить его

в буфере, и многие организации хранят свои системы классификации таким образом. Также характеристики производительности системы хранения данных не зависят от классификатора. Система хранения данных может обрабатывать данные с высокой степенью изменчивости.

Почему системы классификации не так хороши?

Большинство систем хранения данных не имеют хорошего классификатора, а система классификации данных вряд ли со временем станет лучше. Если в вашей системе хранения данных нет хорошего классификатора, у вашей системы классификации возникнут проблемы.

Большинство компаний так не думают о своих системах хранения данных. Вместо этого они предполагают, что систему можно исправить. Они видят в этом то, что со временем можно улучшить, основываясь на будущих усилиях по техническому обслуживанию. Это убеждение также позволяет легко исправить некоторые проблемы, возникающие из-за плохих систем хранения данных. Например, система хранения данных, которая не принимает слишком короткие или неупорядоченные данные, со временем может быть улучшена, если к ее исправлению будет привлечено больше людей.

Суммирование

Суммирование – предоставление более компактного представления набора данных, включая визуализацию структуры данных, полезно для решения более простых задач и поиска данных для статистических закономерностей и выводов. Вы часто можете аппроксимировать эту структуру, моделируя структуру с помощью алгоритма, аналогичного линейному моделированию.

Конец ознакомительного фрагмента.