IBM Analytics
Информационная брошюра

Основополагающая методология анализа и обработки данных

IBM

В области анализа и обработки данных стандартной практикой является решение проблем и получение ответов на вопросы. Зачастую специалисты по анализу и обработке данных для прогнозирования возможных вариантов развития событий строят модели или выявляют базовые закономерности, чтобы достичь необходимого понимания. Впоследствии организации могут предпринять определенные действия на основе выявленных сведений, чтобы максимально улучшить результаты.

Для анализа данных и построения моделей в настоящее время имеется широкий выбор стремительно развивающихся технологий. За удивительно короткий срок эти технологии прошли путь от настольных систем до огромных хранилищ данных с массовым параллелизмом и аналитической функциональностью внутри базы данных в реляционных базах данных и Apache Hadoop. Текстовая аналитика по неструктурированным или полуструктурированным данным приобретает все большее значение как способ включения настроений и другой полезной информации из текста в прогнозные модели, что часто приводит к существенным улучшениям качества и точности моделей.

Зарождающиеся подходы к аналитике нацелены на автоматизацию многих этапов построения и применения моделей, делая технологию машинного обучения более доступной для людей, не обладающих высокими математическими способностями. Кроме того, в отличие от подхода «сверху вниз», при котором сначала определяется бизнес-проблема и затем анализируются данные, чтобы найти ее решение, некоторые специалисты по обработке данных могут использовать обратный подход — «снизу вверх». При втором подходе специалист по обработке данных анализирует большие объемы данных, чтобы выяснить, какую бизнес-цель можно порекомендовать исходя из имеющихся данных, и затем ищет решения для этой проблемы. Поскольку при решении большинства проблем используется подход «сверху вниз», то именно он учитывается в методологии, рассматриваемой в данной публикации.

Методология из 10 этапов для анализа и обработки данных с применением различных подходов и технологий

По мере того как функциональные возможности аналитики данных становятся более доступными и превалирующими, специалисты по обработке данных испытывают потребность в основополагающей методологии, на которой могла бы базироваться стратегия их работы независимо от технологий, объемов данных или задействованных подходов (см. рис. 1). Такая методология имеет определенные сходства с признанными методологиями[1-5] интеллектуального анализа данных, однако в ней повышенное внимание уделено таким новым методам анализа и обработки данных, как использование очень больших объемов данных, включение текстовой аналитики в прогнозные модели и автоматизация некоторых процессов.

Методология состоит из 10 этапов, образующих итерационный процесс использования данных для выявления знаний. В контексте общей методологии ключевую роль играет каждый этап.

Что такое методология?

Методология — это общая стратегия, управляющая процессами и действиями в определенной предметной области. Методология не зависит от определенных технологий или инструментов и не является набором методов или предписаний. Для специалиста по обработке и анализу данных методология является скорее платформой, определяющей порядок использования любых методов, процессов и эвристики для получения необходимых ответов или результатов.

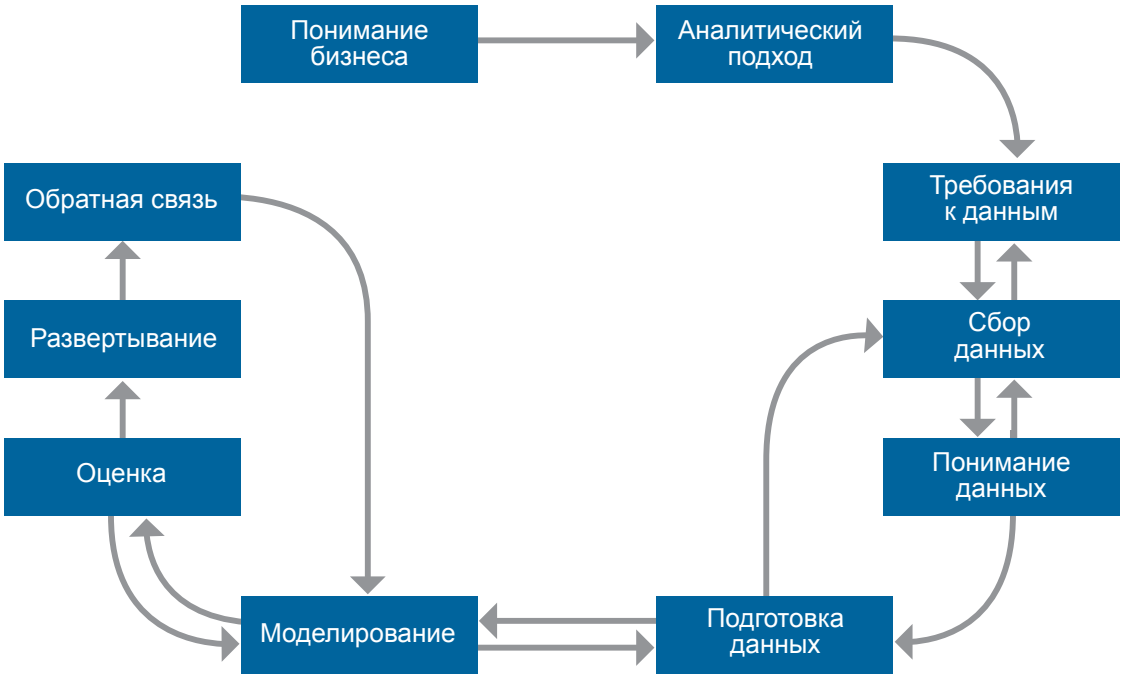

Рисунок 1. Основополагающая методология анализа и обработки данных.

Этап 1: понимание бизнеса

Каждый проект должен начинаться с понимания бизнеса. На данном этапе наиважнейшую роль играют бизнес-заказчики аналитического решения, которые должны определить задачу, цели проекта и требования к решению с точки зрения бизнеса. На первом этапе должен быть заложен фундамент для успешного разрешения бизнес-задачи. Чтобы гарантировать успешную реализацию проекта, бизнес-заказчики должны привлекаться в течение всего проекта и предоставлять экспертные знания в соответствующей области, оценивать промежуточные результаты и ход реализации проекта для создания ожидаемого решения.

Этап 2: аналитический подход

После четкого определения бизнес-задачи специалист по обработке данных может выбрать аналитический подход к ее решению. На данном этапе бизнес-задача должна получить свое выражение в контексте статистических методов и методов машинного обучения, чтобы организация смогла выбрать те методы, которые наилучшим образом подходят для получения ожидаемого результата. Например, если целью является предсказание ответа типа «да» или «нет», то аналитический подход может быть определен как построение, тестирование и внедрение модели классификации.

Этап 3: требования к данным

Требования к данным определяет выбранный аналитический подход. В частности, для использования аналитических методов требуется определенный контент данных, форматы и представления, а также наличие экспертных знаний в предметной области.

Этап 4: сбор данных

На этапе первоначального сбора данных специалисты по анализу и обработке данных идентифицируют и собирают доступные источники структурированных, неструктурированных и полуструктурированных данных, имеющих отношение к предметной области, в которой решается задача. Как правило, на данном этапе принимается решение относительно дополнительных инвестиций в получение менее доступных элементов данных. Наилучший вариант может при этом заключаться в том, чтобы отложить решение о таких инвестициях, пока не появится дополнительная информация о данных и модели. Если собранные данные неполные, специалисту по анализу и обработке данных, возможно, потребуется соответствующим образом пересмотреть требования к данным и заняться сбором новых и/или дополнительных данных.

Несмотря на то что составление выборок данных и формирование поднаборов по-прежнему важны, современные высокопроизводительные платформы и аналитическая функциональность внутри баз данных дают возможность специалистам по анализу и обработке данных использовать намного большие наборы данных, содержащие большую часть доступных данных или даже все. При использовании увеличенного объема данных прогнозные модели могут лучше представлять такие редкие события, как распространенность заболеваний или отказ системы.

Этап 5: понимание данных

После первоначального сбора данных специалисты по их анализу и обработке обычно используют описательную статистику и методы визуализации, чтобы понять контент данных, оценить качество данных и получить из них начальные знания. Для заполнения пробелов в собранных данных может потребоваться сбор дополнительных данных.

Этап 6: подготовка данных

Данный этап охватывает все действия по созданию набора данных, который будет использован на последующем этапе построения модели. Действия по подготовке данных включают очистку данных (обработка отсутствующих или недействительных значений, удаление дубликатов, правильное форматирование), объединение данных из различных источников (файлов, таблиц, платформ) и трансформацию данных в более ценные переменные.

В ходе процесса, который называется *создание новых признаков (feature engineering)*, специалисты по анализу и обработке данных могут создавать дополнительные поясняющие переменные, также называемые *предикторами* или *признаками*, путем комбинирования экспертных знаний в предметной области и существующих структурированных переменных. Если данные доступны в форме текста, например, в виде журналов звонков клиентов в центры телефонного обслуживания или записей врачей в неструктурированной или полуструктурированной форме, текстовая аналитика полезна с точки зрения создания новых структурированных переменных для «обогащения» набора предикторов и улучшения точности модели.

Как правило, подготовка данных является наиболее времяемким этапом в проекте, связанном с обработкой и анализом данных. Во многих предметных областях некоторые этапы подготовки данных являются одинаковыми при решении различных задач. Для ускорения процесса можно использовать заблаговременную автоматизацию некоторых шагов по подготовке данных путем минимизации времени на подготовку нестандартизованными средствами. Используя современные высокопроизводительные массивно-параллельные системы и аналитическую функциональность, выполняемую по месту хранения данных, специалисты по обработке и анализу данных могут более легко и быстро подготавливать данные при применении очень больших наборов данных.

Этап 7: моделирование

Этап моделирования начинается с первой версии подготовленного набора данных и заключается в разработке описательных или прогнозных моделей в соответствии с ранее определенным аналитическим подходом. Для построения прогнозных моделей специалисты по обработке и анализу данных применяют *обучающий*

набор (исторические данные, в которых известен необходимый результат). Процесс моделирования, как правило, отличается большим числом итераций, при этом в результате каждой итерации организации получают промежуточное понимание, ведущее к уточнению требований к подготовке данных и построению модели. Для определенного метода специалисты по обработке и анализу данных могут попробовать различные алгоритмы с соответствующими параметрами, чтобы найти наилучшую модель для доступных переменных.

Этап 8: оценка

Во время разработки модели и до ее развертывания специалист по обработке и анализу данных оценивает качество созданной модели, чтобы убедиться в том, что она полностью подходит для решения поставленной бизнес-задачи. Оценка модели подразумевает вычисление различных диагностических показателей и вывод других данных в виде таблиц и графиков, опираясь на которые специалист по обработке и анализу данных может оценить качество модели и эффективность ее применения для решения поставленной задачи. Для прогнозной модели специалисты по обработке и анализу данных используют тестовый набор, который не зависит от обучающего набора, однако имеет то же распределение вероятностей и известный результат. Тестовый набор используется для оценки модели и при необходимости ее дальнейшего совершенствования. Иногда для окончательной оценки качества к валидационному набору применяется финальная модель.

Кроме того, для дополнительного подтверждения качества модели специалисты по обработке и анализу данных могут применять к ней проверки статистической значимости. Такое дополнительное подтверждение качества может быть полезно при обосновании необходимости внедрить модель или для выполнения действий в условиях повышенной критичности, например, в случае дорогостоящего дополнительного медицинского протокола или критически важной системы управления полетами.

Этап 9: развертывание

После того как подходящая модель будет разработана и утверждена бизнес-заказчиками, выполняется ее развертывание в производственной среде или сравнимой с ней по характеристикам тестовой среде. До установления точных рабочих характеристик модели ее развертывание обычно выполняют в ограниченных условиях. Процесс развертывания может быть простым и заключаться в составлении отчета с рекомендациями или сложным, например, в случае встраивания модели в комплексный рабочий алгоритм и процесс оценивания, управление которым осуществляется с помощью специально созданного приложения. Развертывание модели в рамках операционного бизнес-процесса обычно проводится с привлечением дополнительных групп специалистов, экспертов и технологий, предоставляемых организацией. Например, группа специалистов по продажам может развернуть модель прогнозирования ответов посредством процесса управления рекламными кампаниями, созданного отделом разработки и администрируемого силами отдела маркетинга.

Этап 10: обратная связь

Собирая результаты, предоставляемые внедренной моделью, организация получает обратную связь относительно рабочих характеристик модели и ее влияния на среду, в которой она была развернута. Например, обратная связь может быть в форме показателей отклика на рекламную кампанию, нацеленную на группу клиентов, которые были идентифицированы моделью как лица, которые отреагируют на такую кампанию с высокой вероятностью. Анализ обратной связи дает возможность специалистам по обработке и анализу данных совершенствовать модель для повышения ее точности и полезности. Для ускорения процесса обновления модели и получения улучшенных результатов можно автоматизировать некоторые или все этапы, связанные со сбором сведений обратной связи, оценкой, совершенствованием и повторным развертыванием модели.

Обеспечение непрерывной пользы для организации

Этапы методологии иллюстрируют итерационную природу процесса решения задач. По мере того как специалисты по обработке и анализу данных узнают больше о данных и построении моделей, они часто возвращаются на предыдущий этап для внесения изменений. Создание и развертывание модели не являются единовременными процессами; созданная и развернутая модель постоянно изменяется, улучшается и адаптируется к изменяющимся условиям в ходе получения обратной связи, совершенствования и повторного развертывания. Таким образом, и модель, и усилия по ее созданию могут непрерывно приносить организации пользу,

пока необходимо решение, ради которого создавалась модель.

Более подробная информация

Новый курс по основополагающей методологии обработки и анализа данных доступен в университете больших данных. Бесплатный курс доступен онлайн по следующим ссылкам: http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

Перейдите по следующим ссылкам, чтобы просмотреть рабочие примеры внедрения этой методологии в реальных условиях:

- http://ibm.co/1SUhxFm
- http://ibm.co/1IazTvG

Благодарности

Автор выражает благодарность Майклу Хейду (Michael Haide), Майклу Версту (Michael Wurst, Ph.D.), Брэндону Маккензи (Brandon MacKenzie) и Грегори Родду (Gregory Rodd) за полезные комментарии, а также Джо А. Рамосу (Jo A. Ramos) за его роль в разработке этой методологии в течение многих лет сотрудничества.

Об авторе

Джон Б. Роллинс (John B. Rollins, Ph.D.) работает в организации IBM® Analytics в качестве специалиста по обработке и анализу данных. Г-н Роллинс ранее занимался вопросами инженерных разработок, интеллектуального анализа данных и эконометрики в различных отраслях. Он является владельцем семи патентов и автором популярного учебника по инженерному делу, а также многочисленных публикаций на технические темы. Г-н Роллинс получил степень доктора наук по экономике и нефтегазовому делу в университете A&M штата Техас.

IBM Восточная Европа/Азия
123317, Москва
Краснопресненская наб., 18
Тел.: +7 (495) 775-8800, +7 (495) 940-2000
Факс: +7 (495) 940-2070

[1] Brachman, R. & Anand, T., "The process of knowledge discovery in databases," in Fayyad, U. et al., eds., Advances in knowledge discovery and data mining, AAAI Press, 1996 (pp. 37-57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, "Cross Industry Standard Process for Data Mining," http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E. and Chodagam, J.: Dynamic Warehousing: Data Mining Made Easy, IBM Redbook SG24-7418-00 (Sep. 2007), pp. 9-26.

[5] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, Oct. 28, 2014, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

IMW14828-RURU-00