Введение

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших в последние годы проблем. Главная причина — концепция усреднения по выборке, приводящая к операциям над фиктивными величинами. Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез и для “грубого” разведочного анализа, составляющего основу оперативной аналитической обработки данных.

В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки потоки сырых данных образуют никому не нужную свалку.

Основное различие между статистикой и Data Mining заключается в эффективности алгоритмов и технологичности их применения. Подавляющее большинство классических процедур имеют время выполнения, квадратичное или даже кубическое по объёму исходных данных. При количестве объектов, превосходящем несколько десятков тысяч, они работают неприемлемо медленно даже на самых современных компьютерах. За последние десятилетия значительные усилия в области Data Mining были направлены на создание специализированных алгоритмов, способных выполнять те же задачи за линейное или даже логарифмическое время без существенной потери точности.

1. Определение

Интеллектуальный анализ данных (англ. Data Mining) — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Собственно суть работы Data Mining в преобразовании разрозненных данных в информацию.

2. Задачи решаемые data mining

2.1. Классификация

Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Методы решения. Для решения задачи классификации могут использоваться методы: ближайшего соседа; k-ближайшего соседа; байесовские сети; индукция деревьев решений; нейронные сети.

2.2. Кластеризация

Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.

2.3. Ассоциация

В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Например, поиск «устойчивых связей в корзине покупателя»— вместе с пивом часто покупают орешки.

Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

2.4. Прогнозирование

В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

2.5. Определение отклонений или выбросов

Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

2.6. Оценивание

Задача оценивания сводится к предсказанию непрерывных значений признака.

2.7. Анализ связей

Задача нахождения зависимостей в наборе данных.

2.8. Визуализация

В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

3. Методики Data Mining

В основу современной технологии Data Mining положена концепция шаблонов, 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.

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные регулярности в данных, составляющие так называемые скрытые знания. К обществу пришло понимание, что сырые данные содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие

самородки.

Можно выделить типичный ряд этапов решения задач методами ИАД:

1. Выявление закономерностей (свободный поиск).
2. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование).
3. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

3.1. Свободный поиск

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей. Предварительные гипотезы относительно вида закономерностей здесь не определяются.

Закономерность - существенная и постоянно повторяющаяся взаимосвязь, определяющая этапы и формы процесса становления, развития различных явлений или процессов.

Система Data Mining на этой стадии определяет шаблоны, при использовании других средств нужно формировать запросы. Здесь же аналитик освобождается от такой работы - шаблоны ищет за него система. Особенно полезно применение данного подхода в сверхбольших базах данных, где уловить закономерность путем создания запросов достаточно сложно, для этого требуется перепробовать множество разнообразных вариантов.

Свободный поиск представлен такими действиями:

- выявление закономерностей условной логики;
- выявление закономерностей ассоциативной логики ;
- выявление трендов и колебаний.

3.2. Прогностическое моделирование

Вторая стадия Data Mining - прогностическое моделирование - использует результаты работы первой стадии. Здесь обнаруженные закономерности используются непосредственно для прогнозирования.

Прогностическое моделирование включает такие действия:

- предсказание неизвестных значений;

- прогнозирование развития процессов.

В процессе прогностического моделирования решаются задачи классификации и прогнозирования.

При решении задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью, к одному из известных, предопределенных классов на основании известных значений.

При решении задачи прогнозирования результаты первой стадии (определение тренда или колебаний) используются для предсказания неизвестных (пропущенных или же будущих) значений целевой переменной (переменных).

3.3. Анализ исключений

На третьей стадии Data Mining анализируются исключения или аномалии, выявленные в найденных закономерностях.

Действие, выполняемое на этой стадии, - выявление отклонений . Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии свободного поиска.

4. Методы и основные классы систем Data Mining

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера. Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта. В частности они обладают способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений.

Естественно, разработка таких методов требует значительных интеллектуальных усилий, и нетривиальной автоматизации.

Большинство инструментов Data Mining, реализуют сразу несколько методов, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, самоорганизующиеся карты Кохонена и визуализацию.

В универсальных прикладных статистических пакетах реализуется широкий спектр разнообразнейших методов (как статистических, так и кибернетических). Следует

учитывать, что для возможности их использования, а также для интерпретации результатов работы статистических методов (корреляционного, регрессионного, факторного, дисперсионного анализа и др.) требуются специальные знания в области статистики.

Универсальность того или иного инструмента часто накладывает определенные ограничения на его возможности. Преимуществом использования таких универсальных пакетов является возможность относительно легко сравнивать результаты построенных моделей, полученные различными методами. Такая возможность реализована, например, в пакете Statistica, где сравнение основано на так называемой "конкурентной оценке моделей". Эта оценка состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик для выбора наилучшей из них.

4.1. Нейронные сети

Это большой класс систем, архитектура которых имеет аналогию (как теперь известно, довольно слабую) с построением нервной ткани из нейронов. В одной из наиболее распространенных архитектур, многослойном перцептроне с обратным распространением ошибки, имитируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т. д. Эти значения рассматриваются как сигналы, передающиеся в следующий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. В результате на выходе нейрона самого верхнего слоя вырабатывается некоторое значение, которое рассматривается как ответ — реакция всей сети на введенные значения входных параметров. Для того чтобы сеть можно было применять в дальнейшем, ее прежде надо "натренировать" на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы на них. Тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам.

Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существенный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой “черный ящик”. Знания, зафиксированные как веса нескольких сотен

межнейронных связей, совершенно не поддаются анализу и интерпретации.

4.2. Системы рассуждений на основе аналогичных случаев

Идея систем case based reasoning — CBR — на первый взгляд крайне проста. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы находят в прошлом близкие аналоги наличной ситуации и выбирают тот же ответ, который был для них правильным. Поэтому этот метод еще называют методом "ближайшего соседа". В последнее время распространение получил также термин memory based reasoning, который акцентирует внимание, что решение принимается на основании всей информации, накопленной в памяти.

Системы CBR показывают неплохие результаты в самых разнообразных задачах. Главным их минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт, — в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов CBR системы строят свои ответы.

Другой минус заключается в произволе, который допускают системы CBR при выборе меры "близости". От этой меры самым решительным образом зависит объем множества прецедентов, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза.

4.3. Деревья решений

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа "ЕСЛИ... ТО..." (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид "значение параметра А больше х?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный — то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить “лучшие” (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и “цепляют” фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода.

4.4. Эволюционное программирование

Проиллюстрируем современное состояние данного подхода на примере системы PolyAnalyst — отечественной разработке, получившей сегодня общее признание на рынке Data Mining. В данной системе гипотезы о виде зависимости целевой переменной от других переменных формулируются в виде программ на некотором внутреннем языке программирования. Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Когда система находит программу, более или менее удовлетворительно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных дочерних программ те, которые повышают точность. Таким образом система "выращивает" несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости. Специальный модуль системы PolyAnalyst переводит найденные зависимости с внутреннего языка системы на понятный пользователю язык (математические формулы, таблицы и пр.).

Другое направление эволюционного программирования связано с поиском зависимости целевых переменных от остальных в форме функций какого-то определенного вида. Например, в одном из наиболее удачных алгоритмов этого типа — методе группового учета аргументов (МГУА) зависимость ищут в форме полиномов. В настоящее время из продающихся в России систем МГУА реализован в системе NeuroShell компании Ward Systems Group.

4.5. Генетические алгоритмы

Data Mining не основная область применения генетических алгоритмов. Их нужно рассматривать скорее как мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее генетические алгоритмы вошли сейчас в стандартный инструментарий методов Data Mining, поэтому они и включены в данный обзор.

Первый шаг при построении генетических алгоритмов — это кодировка исходных логических закономерностей в базе данных, которые именуют хромосомами, а весь набор таких закономерностей называют популяцией хромосом. Далее для реализации концепции отбора вводится способ сопоставления различных хромосом. Популяция обрабатывается с помощью процедур репродукции, изменчивости (мутаций), генетической композиции. Эти процедуры имитируют биологические процессы. Наиболее важные среди них: случайные мутации данных

в индивидуальных хромосомах, переходы (кроссинговер) и рекомбинация генетического материала, содержащегося в индивидуальных родительских хромосомах (аналогично гетеросексуальной репродукции), и миграции генов. В ходе работы процедур на каждой стадии эволюции получаются популяции со все более совершенными индивидуумами.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.

Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения “лучшего” решения. Как и в реальной жизни, эволюцию может “заклинить” на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

4.6. Алгоритмы ограниченного перебора

Алгоритмы ограниченного перебора были предложены в середине 60-х годов М.М. Бонгардом для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.

Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: $X = a$; $X < a$; $X > a$; $a < X < b$ и др., где X — какой либо параметр, “a” и “b” — константы. Ограничением служит длина комбинации простых логических событий. На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.

Основной недостаток — системы, основанные на таких алгоритмах как правило выдают решение за приемлемое время только для сравнительно небольшой размерности данных.

4.7. Системы для визуализации многомерных данных

В той или иной мере средства для графического отображения данных поддерживаются всеми системами Data Mining. Вместе с тем, весьма внушительную долю рынка занимают системы, специализирующиеся исключительно на этой функции. В подобных системах основное внимание сконцентрировано на дружелюбности пользовательского интерфейса, позволяющего ассоциировать с анализируемыми показателями различные параметры диаграммы рассеивания объектов (записей) базы данных. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Кроме того, системы визуализации данных снабжены удобными средствами для масштабирования и вращения изображений.

5. Сфера применения

Сфера применения Data Mining ничем не ограничена — она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня, мягко говоря, заинтриговали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных. Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, известны сообщения об экономическом эффекте, в 10–70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. Известны сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример — годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании.

Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

6. Некоторые бизнес-приложения Data Mining

6.1. Розничная торговля

Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля. Вот типичные задачи, которые можно решать с помощью Data Mining в сфере розничной торговли:

- анализ покупательской корзины (анализ сходства) предназначен для выявления товаров, которые покупатели стремятся приобретать вместе. Знание покупательской корзины необходимо для улучшения рекламы,

выработки стратегии создания запасов товаров и способов их раскладки в торговых залах.
- исследование временных шаблонов помогает торговым предприятиям принимать решения о создании товарных запасов. Оно дает ответы на вопросы типа "Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?"
- создание прогнозирующих моделей дает возможность торговым предприятиям узнавать характер потребностей различных категорий клиентов с определенным поведением, например, покупающих товары известных дизайнеров или посещающих распродажи. Эти знания нужны для разработки точно направленных, экономичных мероприятий по продвижению товаров.

6.2. Банковское дело

Достижения технологии Data Mining используются в банковском деле для решения следующих распространенных задач:

- выявление мошенничества с кредитными карточками. Путем анализа прошлых транзакций, которые впоследствии оказались мошенническими, банк выявляет некоторые стереотипы такого мошенничества.
- сегментация клиентов. Разбивая клиентов на различные категории, банки делают свою маркетинговую политику более целенаправленной и результативной, предлагая различные виды услуг разным группам клиентов.
- прогнозирование изменений клиентуры. Data Mining помогает банкам строить прогнозные модели ценности своих клиентов, и соответствующим образом обслуживать каждую категорию.

6.3. Телекоммуникации

В области телекоммуникаций методы Data Mining помогают компаниям более энергично продвигать свои программы маркетинга и ценообразования, чтобы удерживать существующих клиентов и привлекать новых. Среди типичных мероприятий отметим следующие:

- анализ записей о подробных характеристиках вызовов. Назначение такого анализа — выявление категорий клиентов с похожими стереотипами пользования их услугами и разработка привлекательных наборов цен и услуг;
- выявление лояльности клиентов. Data Mining можно использовать для определения характеристик клиентов, которые, один раз воспользовавшись услугами данной компании, с большой долей вероятности останутся ей

верными. В итоге средства, выделяемые на маркетинг, можно тратить там, где отдача больше всего.

6.4. Страхование

Страховые компании в течение ряда лет накапливают большие объемы данных. Здесь обширное поле деятельности для методов Data Mining:

- выявление мошенничества. Страховые компании могут снизить уровень мошенничества, отыскивая определенные стереотипы в заявлениях о выплате страхового возмещения, характеризующих взаимоотношения между юристами, врачами и заявителями.
- анализ риска. Путем выявления сочетаний факторов, связанных с оплаченными заявлениями, страховщики могут уменьшить свои потери по обязательствам. Известен случай, когда в США крупная страховая компания обнаружила, что суммы, выплаченные по заявлениям людей, состоящих в браке, вдвое превышает суммы по заявлениям одиноких людей. Компания отреагировала на это новое знание пересмотром своей общей политики предоставления скидок семейным клиентам.

Можно привести еще много примеров различных областей знания, где методы Data Mining играют ведущую роль. Особенность этих областей заключается в их сложной системной организации. Они относятся главным образом к надкибернетическому уровню организации систем, закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей. Данные в указанных областях неоднородны, гетерогенны, нестационарны и часто отличаются высокой размерностью.

Заключение

Главная ценность Data Mining - это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого можно принимать решения.

Рынок систем Data Mining экспоненциально развивается. В этом развитии принимают участие практически все крупнейшие корпорации. В частности Microsoft непосредственно руководит большим сектором данного рынка (издает специальный журнал, проводит конференции, разрабатывает собственные продукты).

Системы Data Mining применяются по двум основным направлениям: 1) как массовый продукт для бизнес-приложений; 2) как инструменты для проведения уникальных исследований (генетика, химия, медицина и пр.). Количество инсталляций массовых продуктов, судя по имеющимся сведениям, сегодня достигает десятков тысяч. Лидеры Data Mining связывают будущее этих систем с использованием их в качестве интеллектуальных приложений, встроенных в корпоративные хранилища данных.

Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону логических алгоритмов поиска в данных if-then правил. С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации БД, извлечения из данных “скрытых” знаний, интерпретации данных, установления ассоциаций в БД и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

Вместе с тем, главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор, либо строят деревья решений, имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил.