УДК 004.82

Г. Г. АСЕЕВ, д-р. техн. наук, профессор (г. Харьков)

ПРОБЛЕМА ОБНАРУЖЕНИЯ НОВОГО ЗНАНИЯ В ХРАНИЛИЩАХ ДАННЫХ МЕТОДАМИ KNOWLEDGE DISCOVERY IN DATABASES

Стаття присвячена проблемі виявлення нового знання в сховищах даних – Knowledge Discovery in Databases – (KDD) і основному крокові цього процесу – Data Mining (DM) (розробці даних або, дослівно, «видобуткові даних»).

К настоящему времени сложилась ситуация, что происходит подмена понятий «обнаружение знаний в базах данных» методами Knowledge Discovery in Databases – (KDD) и «добыча данных» или «разработка данных» методами Data Mining (DM) [1-3]. Эта статья посвящена проблеме обнаружения нового знания в хранилищах данных – (KDD – Knowledge Discovery in Databases), а основной шаг этого процесса – Data Mining (DM) (разработка данных или, дословно, «добыча данных»), будет подвергнут анализу в следующей публикации.

Что такое Knowledge Discovery in Databases? Knowledge Discovery in Databases – аналитический процесс исследования человеком большого объема информации с привлечением средств автоматизированного исследования данных с целью обнаружения скрытых в данных структур или зависимостей. Предполагается полное или частичное отсутствие априорных представлений о характере скрытых структур и зависимостей. KDD включает предварительное осмысление и неполную формулировку задачи (в терминах целевых переменных), преобразование данных к доступному для автоматизированного анализа формату и их предварительную обработку, обнаружение средствами автоматического исследования данных (Data Mining) скрытых структур или зависимостей, апробация обнаруженных моделей на новых, не использовавшихся для построения моделей данных и интерпретация человеком обнаруженных моделей. KDD включает в себя проблемы использования источников данных, хранения данных, подготовки исходного набора данных, предобработки и очистки исходных данных, трансформации и нормализации данных, выдвижения гипотез, применения методов DM (построение моделей анализа, использование модели, наблюдение за моделью, постобработка данных и интерпретация полученных результатов, вывода систем отчетности).

Очертим некоторые рекомендации, следуя которым, можно подготовить качественные данные в нужном объеме для анализа. В этой последовательности действий все достаточно просто и логично, но, несмотря на это, пользователи почти всегда допускают одни и те же тривиальные ошибки. Надеемся, что изложенная последовательность действий позволит

допускать меньше ошибок такого рода. Ниже описан не жесткий набор правил, а, скорее, список рекомендаций, которых желательно придерживаться.

Источники данных. В качестве первичного источника данных должны выступать все сведения, которые могут пригодиться для принятия решения: базы данных систем управления предприятием, офисные документы, Интернет. Причем речь идет не только о внутренних, но и о внешних данных (макроэкономические показатели, конкурентная среда, демографические показатели и т. п.).

Хранение данных. Хотя в хранилище данных не реализуются технологии анализа, оно является той базой, на которой нужно строить аналитическую систему. При отсутствии хранилища на сбор и систематизацию необходимой для анализа информации будет уходить большая часть времени, что в значительной степени сведет на нет все достоинства анализа - ведь одним из ключевых показателей любой аналитической системы является возможность быстро получить результат.

Подготовка исходного набора данных. Этот этап заключается в создании набора данных, в том числе из различных источников, выбора обучающей выборки и т. д. Для этого должны существовать развитые инструменты доступа к различным источникам данных. Желательна поддержка работы с хранилищами данных и наличие семантического слоя, позволяющего использовать для подготовки исходных данных не технические термины, а бизнес понятия.

Методика анализа с использованием механизмов Data Mining базируется на различных алгоритмах извлечения закономерностей из исходных данных, результатом работы которых являются модели. Таких алгоритмов довольно много, но, несмотря на их обилие, использование машинного обучения и т. п., они не способны гарантировать качественное решение. Никакой самый изощренный метод сам по себе не даст хороший результат, т. к. критически важным становится вопрос качества исходных данных. Чаще всего именно качество данных является причиной неудачи.

Для того чтобы найти новое знание на основе данных большого хранилища недостаточно просто взять алгоритмы Data Mining, запустить их и ждать появления интересных результатов. Нахождение нового знания – это процесс, который включает в себя несколько шагов, каждый из которых необходим для уверенности в эффективном применении средств Data Mining.

В процессе подбора влияющих факторов необходимо максимально абстрагироваться от информационных систем и имеющихся в наличии данных. Очень часто встречается ситуация, когда пользователи говорят: «Вот есть такие данные, что можно на них получить?». Это порочная практика – мы должны решать задачу и подбирать данные для ее решения, а не брать имеющуюся информацию и придумывать что из них можно «выжать». Целью является решение актуальной задачи, а не оправдание затрат на сбор большого объема данных.

После подготовки таблицы с описанием факторов нужно экспертно оценить значимость каждого из факторов. Эта оценка не является окончательной, она будет отправной точкой. В процессе анализа вполне может оказаться, что фактор, который эксперты посчитали очень важным, таковым, по сути, не является и, наоборот, незначимый с их точки зрения фактор может оказывать значительное влияние. Но в любом случае, все варианты проанализировать сразу невозможно, нужно от чего–то отталкиваться, этой точкой и является оценка экспертов. К тому же, довольно часто реальные данные подтверждают их оценку.

Формализация и сбор данных. Далее необходимо определить способ представления данных, выбрав один из 4–х видов – число, строка, дата, логическая переменная (да/нет). Установить способ представления, т.е. формализовать некоторые данные просто – например, объем продаж в грн, это определенное число. Но довольно часто возникают ситуация, когда непонятно как представить фактор. Чаще всего такие проблемы возникают с качественными характеристиками. Например, на объемы продаж влияет качество товара. Качество – это довольно сложное понятие, но если этот показатель действительно важен, то нужно придумать способ его формализации. Например, определять качество по количеству брака на тысячу единиц продукции, либо экспертно оценивать, разбив на несколько категорий – отлично/хорошо/удовлетворительно/плохо.

Необходимо оценить стоимость сбора нужных для анализа данных. Дело в том, что некоторые данные легко доступны, например, их можно извлечь из существующих информационных систем. Но есть информация, которую не просто собрать, например, сведения о конкурентах. Поэтому необходимо оценить, во что обойдется сбор данных.

Чем больше будет данных для анализа, тем лучше, отбросить их можно на следующих этапах работ – это проще, чем собрать новые сведения. К тому же, необходимо учитывать, что не всегда экспертная оценка значимости факторов будет совпадать с реальной. Т.е. в начале не известно, что на самом деле является значимым, а что нет. Мы отталкиваемся от мнения экспертов относительно значимости факторов, но в действительности все может быть иначе. Поэтому желательно иметь большее число данных, чтобы иметь возможность оценить влияние максимального количества показателей.

Но сбор данных не является самоцелью. Если информацию получить легко, то, естественно, нужно ее собрать. Если данные получить сложно, то необходимо соизмерить затраты на ее сбор и систематизацию с ожидаемыми результатами. Есть несколько методов сбора, необходимых для анализа данных [4, 5].

Получение из учетных систем. Обычно, в учетных системах есть различные механизмы построения отчетов и экспорта данных, поэтому извлечение нужной информации их них, чаще всего, относительно несложная операция.

Получение сведений из косвенных данных. О многих показателях можно

судить по косвенным признакам и этим нужно воспользоваться. Например, можно оценить реальное финансовое положение жителей определенного региона следующим образом. В большинстве случаев имеется несколько товаров, предназначенных для выполнения одной и той же функции, но отличающихся по цене: товары для бедных, средних и богатых. Если получить отчет о продажах товара в интересующий регион и проанализировать пропорции, в которых продаются товары для бедных, средних и богатых, то можно предположить, что чем больше доля дорогих изделий из одной товарной группы, тем более состоятельны в среднем жители данного региона.

Использование открытых источников. Большое количество данных присутствует в открытых источниках, таких как статистические сборники, отчеты корпораций, опубликованные результаты маркетинговых исследований и прочее.

Проведение собственных маркетинговых исследований и аналогичных мероприятий по сбору данных. Это может быть достаточно дорогостоящим мероприятием, но, в любом случае, такой вариант сбора данных возможен.

Ввод данных «вручную», когда данные вводится по различного рода экспертным оценкам сотрудниками организации. Этот метод наиболее трудоемкий.

Стоимость сбора информации различными методами существенно отличается по цене и необходимому для этого времени, поэтому нужно соизмерять затраты с результатами. Возможно, от сбора некоторых данных придется отказаться, но факторы, которые эксперты оценили, как наиболее значимые нужно собрать обязательно, не смотря на стоимость этих работ, либо вообще отказаться от анализа. Очевидно, что если эксперт указал на некоторый фактор как важный, то не учитывать его просто нельзя, т. к. мы рискуем провести анализ, ориентируясь на второстепенные малозначащие факторы. И, следовательно, получить модель, которая будет давать плохие и нестабильные результаты. А такая модель не представляет практической ценности.

Собранные данные нужно преобразовать к единому формату, например, Excel, текстовой файл с разделителями, либо любая СУБД. Данные обязательно должны быть унифицированы, т.е. одна и та же информация везде должна описываться одинаково. Обычно проблемы с унификацией возникают при сборе информации из разнородных источников. В этом случае унификация является серьезной задачей.

При построении модели необходимо помнить одно правило, касающееся корректности исходных данных: если на вход задачи поступает «мусор», то и результатом тоже будет «мусор». Исходные данные могут находиться или в одной базе, или в нескольких. Перед «загрузкой» данных в хранилище необходимо учесть, что различные источники данных могут быть спроектированы под определенные задачи и, соответственно, возникает проблемы, связанные с объединением данных: различные форматы

представления данных (одна и та же по смыслу переменная – например, количество – может быть представлена в различных базах разными типами данных – int или short); разное кодирование данных (например, разный формат даты); различные способы хранения данных; отличающиеся единицы измерения (дюймы и сантиметры); а также частота сбора данных и дата последнего обновления. Даже если данные находятся в одной базе, то все равно надо обращать пристальное внимание на пропущенные значения и значения, нарушающие целостность базы («выбросы»).

Аналитик должен всегда знать, как, где и при каких условиях собираются данные, и быть уверенным, что все данные, которые используются для проведения анализа, измеряют одно и то же одинаковым способом.

Предобработка и очистка исходных данных. Для того, чтобы эффективно применять методы Data Mining, следует обратить серьезное внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т. д. Кроме того, данные могут быть избыточны, недостаточны и т. д. В некоторых задачах требуется дополнить данные некоторой априорной информацией. Наивно предполагать, что если подать данные на вход системы в существующем виде, то на выходе получим полезные знания. Данные должны быть качественны и корректны с точки зрения используемого метода DM. Поэтому один из основных этапов KDD заключается в предобработке данных. Более того, иногда размерность исходного пространства может быть очень большой, и тогда желательно применение специальных алгоритмов понижения размерности. Это как отбор значимых признаков, так и отображение данных в пространство меньшей размерности.

При больших объемах хранилищ данных практически невозможно напрямую применить методы DM. Если не будут произведены формализация и сбор данных и их представления, как описано в предыдущих подразделах, то тогда плохое «качество» исходных данных явится одной из самых серьезных проблем. В связи с тем, что в большинстве случаев источником информации для аналитических систем является корпоративное хранилище данных, в котором аккумулируются сведения из множества разнородных источников, острота проблемы существенно возрастает.

Необходимость предварительной обработки при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. При использовании же механизмов анализа, в основе которых лежат самообучающиеся алгоритмы, такие как нейронные сети, деревья решений и прочее, хорошее качество данных является ключевым требованием.

Очевидно, что исходные («сырые») данные чаще всего нуждаются в очистке. В процессе этого восстанавливаются пропущенные данные, редактируются аномальные значения, вычитается шум, проводится сглаживание. При этом используются алгоритмы робастной фильтрации,

спектрального и вейвлет анализа, последовательной рекуррентной фильтрации, статистического анализа.

На этом этапе производится построение витрины данных, которая будет подвергаться дальнейшей обработке, или, говоря другими словами, производится наполнение витрины или «загрузка» в неё тех данных, которые были отобраны на предыдущих этапах. В это же время производится исправление всех ошибок, которые были выявлены, то есть очистка. Существуют различные аспекты очистки данных. Все они направлены на нахождение и исправление ошибок, которые были допущены ещё на этапе сбора информации. Ошибкой в данных могут считаться:

• пропущенное значение;

• невозможное событие (неверно набранное значение – «выброс»).

Коррекция производится на основе здравого смысла, использования правил и/или с привлечением хорошо знающего предметную область эксперта. То есть транзакция или запись в базе данных, в которой есть такая ошибка, может быть исправлена или, в спорных случаях, исключена из дальнейшего рассмотрения.

После проверки согласованности данных, данные преобразовываются и переформатируются в соответствии с результатами оценки. Это делается для большего удобства наблюдения за данными. Данные дискретных событий преобразовываются в специально разработанную или стандартную форму, если таковая имеется, в которой отражаются время и описание событий. Когда пользователи будут легко разбираться в этой форме, они смогут быстро изучить события, которые лежали в основе построения этой формы. Может показаться, что этот шаг дублирует этап сбора данных, но на самом деле это два совершенно разных этапа. На первом из них происходит отбор данных для ускорения машинной обработки информации (анализа) без потери качества, на втором данные приводятся к виду, удобному для визуального контроля пользователя. Теперь человек проводящий анализ может наиболее полно представить себе исходные данные. Это бывает необходимо для различного рода отчетов, когда необходимо кратко охарактеризовать исходные данные применяемые для анализа.

Для анализируемых процессов различной природы, данные должны быть подготовлены специальным образом для упорядоченных данных, неупорядоченных данных и транзакционных данных.

Упорядоченные данные. Такие данные нужны для решения задач прогнозирования, когда необходимо определить каким образом поведет себя тот или иной процесс в будущем на основе имеющихся исторических данных. Чаще всего в качестве одного из фактов выступает дата или время, хотя это и не обязательно, речь может идти и о неких отсчетах, например, данные, с определенной периодичностью собираемые с датчиков.

Для упорядоченных данных (обычно это временные ряды), каждому столбцу соответствует один фактор, а в каждую строку заносятся упорядоченные по времени события с единым интервалом между строками.

Не допускается наличие группировок, итогов и прочее – нужна обычная таблица.

Если для процесса характерна сезонность/цикличность, необходимо иметь данные хотя бы за один полный сезон/цикл с возможностью варьирования интервалов (понедельное, помесячное…). Т. к. цикличность может быть сложной, например, внутри годового цикла квартальные, а внутри кварталов недельные, то необходимо иметь полные данные как минимум за один самый длительный цикл.

Максимальный горизонт прогнозирования зависит от объема данных:

• данные на 1,5 года – прогноз максимум на 1 месяц;

• данные за 2–3 года – прогноз максимум на 2 месяца.

В общем случае максимальный горизонт прогнозирования (время, на которое можно строить достаточно достоверные прогнозы) ограничивается не только объемом данных. Мы исходим из предположения, что факторы, определяющие развитие процесса будут оказывать влияние и в будущем примерно такое же, что и на текущий момент. Данное предположение справедливо не всегда. Например, в случае слишком быстрого изменения ситуации, появления новых значимых факторов и т.п. это правило не работает. Поэтому в зависимости от задачи требования к объему могут сильно изменяться. Использование слишком большого объема данных для анализа так же нецелесообразно, т. к. в этом случае мы будем строить модель по старой истории, и, следовательно, возможно будем учитывать факторы, возможно, уже утратившие свою значимость.

Неупорядоченные данные. Такого рода данные нужны для задач, где временной фактор не имеет значения, например, оценка кредитоспособности, диагностика, сегментация потребителей. В таких случаях мы считаем ситуацию статичной и поэтому информация о том, что одно событие произошло раньше другого, значения не имеет.

Для неупорядоченных данных каждому столбцу соответствует фактор, а в каждую строку заносится пример (ситуация, прецедент). Упорядоченность строк не требуется. Не допускается наличие группировок, итогов и прочее – нужна обычная таблица.

Количество примеров (прецедентов) должно быть значительно больше количества факторов. В противном случае высока вероятность, что случайный фактор окажет серьезное влияние на результат. Если нет возможности увеличить количество данных, то придется уменьшить количество анализируемых факторов, оставив наиболее значимые.

Желательно, чтобы данные покрывали как можно больше ситуаций реального процесса и пропорции различных примеров (прецедентов) должны примерно соответствовать реальному процессу. Мы пытаемся построить модели на основе предложенных данных, поэтому, чем ближе данные к действительности, тем лучше. Необходимо понимать, что система не может знать о чем–либо, что находится за пределами собранных для анализа данных. Например, если при создании системы диагностики больных

подавать только сведения о больных, то система не будет знать о существовании в природе здоровых людей. И соответственно, любой человек с ее точки зрения будет обязательно чем–то болен.

Транзакционные данные. Транзакционные данные используются в алгоритмах поиска ассоциативных правил, этот метод часто называют «анализом потребительской корзины». Под транзакцией подразумевается несколько объектов или действий, сгруппированных в логически связанную единицу. Очень часто данный механизм используется для анализа покупок (чеков) в супермаркетах. Но, в общем случае, речь может идти о любых связанных объектах или действиях, например, продажа туристических туров с набором сопутствующих услуг (оформление виз, доставка в аэропорт, услуги гида и прочее). Используя данный метод анализа, находятся зависимости вида, «если произошло событие ***А***, то с определенной вероятностью произойдет событие ***Б***».

Анализ транзакций целесообразно производить на большом объеме данных, иначе могут быть выявлены статистически необоснованные правила. Алгоритмы поиска DM способны быстро перерабатывать огромные массивы информации, т. к. основное достоинство алгоритмов заключается именно в масштабируемости, т.е. способности обрабатывать большие объемы данных. Примерное соотношение между количеством объектов и объемом данных:

• 300–500 объектов – более 10 тыс. транзакций;

• 500–1000 объектов – более 300 тысяч транзакций.

При недостаточном количестве транзакций целесообразно уменьшить количество анализируемых объектов, например, сгруппировав их.

Но термин «предобработка» можно трактовать шире, а именно как процесс предварительного экспресс анализа данных. Например, оценить фактор как значимый или нет, все ли факторы учтены для объяснения поведения результирующей величины и т.д. Для этих целей используются такие алгоритмы, как корреляционный анализ, факторный анализ, метод главных компонент, регрессионный анализ и др. (http://www.basegroup.ru/).

Трансформация, нормализация данных. Этот шаг необходим для тех методов, которые требуют, чтобы исходные данные были в каком-то определенном виде. Необходимо стараться приближать данные к нормальному или равномерному распределению. Часто используют следующий способ нормализации: пусть есть a - среднее и σ - дисперсия. Тогда $x \to \frac{x-a}{\sigma}$. Часто такое преобразование проводят перед подачей данных на вход нейросетей.

Выдвижение гипотез. Для того чтобы наиболее полно использовать все преимущества технологий Data Mining необходимо ясно представить цели будущего анализа. В зависимости от целей проводится построение модели. Гипотезой в данном случае будем считать предположение о влиянии определенных факторов на исследуемую нами задачу. Форма этой

зависимости в данном случае значения не имеет. Т.е. мы может сказать, что на продажи влияет отклонение нашей цены на товар от среднерыночной, но при этом не указывать, как, собственно, этот фактор влияет на продажи. Для решения этой задачи и используется Data Mining. Автоматизировать процесс выдвижения гипотез не представляется возможным, по крайней мере, на сегодняшнем уровне развития технологий. Эту задачу должны решать эксперты – специалисты в предметной области. Полагаться можно и нужно на их опыт и здравый смысл. Нужно постараться максимально использовать их знание о предмете и собрать как можно больше гипотез/предположений. Обычно для этих целей хорошо работает тактика мозгового штурма. На первом шаге нужно собрать и систематизировать все идеи, их оценку будем производить позже. Результатом данного шага должен быть список с описанием всех факторов.

Например, для задачи прогнозирования спроса это может быть список следующего вида: сезон, день недели, объемы продаж за предыдущие недели, объем продаж за аналогичный период прошлого года, рекламная компания, маркетинговые мероприятия, качество продукции, бренд, отклонение цены от среднерыночной, наличие данного товара у конкурентов…

Выводы. Суммируя вышесказанное, видно, что обнаружение новых знаний в больших хранилищах данных не является простым применением методов Data Mining. Иерархически выше находятся процессы исследования большого объема информации с привлечением средств Knowledge Discovery in Databases, которые включают в себя проблемы использования источников данных, хранения данных, подготовки исходного набора данных, предобработки и очистки исходных данных, трансформации и нормализации данных, выдвижения гипотез, которые подготавливают базу применения методов DM (построение моделей анализа, использование модели, наблюдение за моделью, постобработка данных и интерпретация полученных результатов, вывода систем отчетности). Предложенная последовательность действий не зависит от предметной области, поэтому ее можно использовать для любой сферы деятельности.

Список литературы: 1. *Мусаев А. А.* Интеллектуальный анализ данных: Клондайк или Вавилон? [Электронный ресурс] / А. А. Мусаев. – Режим доступа: http://www.bizcom.ru /analisys/ – Загл. с экрана. **2.** *Арсеньев С. Б.* Использование технологии анализа данных в интеллектуальных информационных системах. Управление информационными потоками / С. Б. Арсеньев, В. Б. Бритков, Н. А. Маленкова // Сб. тр. Ин-та систем. анализа РАН. – М.: Эдиториал УРСС, 2002. – С. 47–68. **3.** *Орехов С. В.* Применение технологии Data Mining при решении задачи ситуационного управления системой водоснабжения города в условиях аварийности / С. В. Орехов // Вест. нац. техн. ун-та «ХПИ». - № 6. – 2003. – С. 97-101. **4.** *Арустамов А.* Анализируй это / А. Арустамов // Компьютерра. – 2002. – № 21 (446). – С. 18-24. **5.** *Арустамов А.* Data Mining – подготовка исходных данных [Электронный ресурс] / А. Арустамов. – Режим доступа: http://www.basegroup.ru/tasks/. – Загл. с экрана.

Поступила в редколлегию 07.03.06