ЛЕКЦИЯ 3 МЕТОДЫ И СТАДИИ DATA MINING

Основная особенность *Data Mining* - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических *методов*) и последних достижений в сфере информационных технологий. В технологии *Data Mining* гармонично объединились строго формализованные *методы* и *методы* неформального анализа, т.е. количественный и качественный *анализ* данных.

К *методам* и *алгоритмам Data Mining* относятся следующие: *искусственные нейронные сети*, деревья решений, символьные правила, *методы* ближайшего соседа и k-ближайшего соседа, *метод опорных векторов*, байесовские сети, линейная регрессия, корреляционно-регрессионный *анализ*; иерархические *методы* кластерного анализа, неиерархические *методы* кластерного анализа, в том числе *алгоритмы* k-средних и k-медианы; *методы* поиска ассоциативных правил, в том числе *алгоритм* Apriori; метод ограниченного перебора, *эволюционное программирование* и генетические *алгоритмы*, разнообразные *методы* визуализации данных и множество других *методов*.

Большинство аналитических *методов*, используемые в технологии *Data Mining* - это известные математические *алгоритмы* и *методы*. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств. Следует отметить, что большинство *методов Data Mining* были разработаны в рамках теории искусственного интеллекта.

Метод (*method*) представляет собой норму или правило, определенный *путь*, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Понятие *алгоритма* появилось задолго до создания электронных вычислительных машин. Сейчас *алгоритмы* являются основой для решения многих прикладных и теоретических задач в различных сферах человеческой деятельности, в большинстве - это задачи, решение которых предусмотрено с использованием компьютера.

Алгоритм (*algorithm*) - точное предписание относительно последовательности действий (шагов), преобразующих исходные данные в искомый результат.

Классификация стадий Data Mining

Data Mining может состоять из двух или трех стадий:

Стадия 1. Выявление *закономерностей* (***свободный поиск***).

Стадия 2. Использование выявленных *закономерностей* для предсказания неизвестных значений (***прогностическое моделирование***).

В дополнение к этим стадиям иногда вводят стадию валидации, следующую за стадией *свободного поиска*. Цель валидации - проверка достоверности найденных *закономерностей*. Однако, мы будем считать валидацию частью первой стадии, поскольку в реализации многих *методов*, в частности, нейронных сетей и деревьев решений, предусмотрено деление общего множества данных на обучающее и проверочное, и последнее позволяет проверять достоверность полученных результатов.

Стадия 3. ***Анализ исключений*** - стадия предназначена для выявления и объяснения аномалий, найденных в *закономерностях*.

Итак, процесс *Data Mining* может быть представлен рядом таких последовательных стадий:

СВОБОДНЫЙ ПОИСК (в том числе ВАЛИДАЦИЯ) ->

-> ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ->

-> АНАЛИЗ ИСКЛЮЧЕНИЙ

1. ***Свободный поиск (Discovery)***

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

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

Система *Data Mining* на этой стадии определяет шаблоны, для получения которых в системах OLAP, например, аналитику необходимо обдумывать и создавать множество запросов. 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 (conditional logic);

- выявление *закономерностей* ассоциативной логики (associations and *affinities*);
- выявление трендов и колебаний (*trends* and *variations*).

Допустим, имеется база данных кадрового агентства с данными о профессии, стаже, возрасте и желаемом уровне вознаграждения. В случае самостоятельного задания запросов аналитик может получить приблизительно такие результаты: средний желаемый уровень вознаграждения специалистов в возрасте от 25 до 35 лет равен 1200 условных единиц. В случае *свободного поиска* система сама ищет *закономерности*, необходимо лишь задать целевую переменную. В результате поиска *закономерностей* система сформирует набор логических правил "если ..., то ...".

Могут быть найдены, например, такие *закономерности* " **Если** возраст < 20 лет и желаемый уровень вознаграждения > 700 условных единиц, то в 75% случаев соискатель ищет работу программиста" или " **Если** возраст >35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90% случаев соискатель ищет руководящую работу". Целевой переменной в описанных правилах выступает профессия.

При задании другой целевой переменной, например, возраста, получаем такие правила: " **Если** соискатель ищет руководящую работу и его стаж > 15 лет, то возраст соискателя > 35 лет в 65 % случаев".

Описанные действия, в рамках стадии *свободного поиска*, выполняются при помощи :

- индукции правил условной логики (задачи классификации и кластеризации, описание в компактной форме близких или схожих групп объектов);
- индукции правил ассоциативной логики (задачи ассоциации и последовательности и извлекаемая при их помощи информация);
- определения трендов и колебаний (исходный этап задачи прогнозирования).

На стадии *свободного поиска* также должна осуществляться валидация *закономерностей*, т.е. проверка их достоверности на части данных, которые не принимали участие в формировании *закономерностей*. Такой прием разделения данных на обучающее и проверочное множество часто используется в методах нейронных сетей и деревьев решений и будет описан в соответствующих лекциях.

2. ***Прогностическое моделирование*** (*Predictive Modeling*)

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 (*outcome* prediction);
- прогнозирование развития процессов (*forecasting*).

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).

Продолжая рассмотренный пример первой стадии, можем сделать следующий вывод.

Зная, что соискатель ищет руководящую работу и его стаж > 15 лет, на 65 % можно быть уверенным в том, что возраст соискателя > 35 лет. Или же, если возраст соискателя > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, на 90% можно быть уверенным в том, что соискатель ищет руководящую работу.

Сравнение свободного поиска и прогностического моделирования с точки зрения логики

Свободный поиск раскрывает общие *закономерности*. Он по своей природе индуктивен. Закономерности, полученные на этой стадии, формируются от частного к общему. В результате мы получаем некоторое общее знание о некотором классе объектов на основании исследования отдельных представителей этого класса.

Правило: "Если возраст соискателя < 20 лет и желаемый уровень вознаграждения > 700 условных единиц, то в 75% случаев соискатель ищет работу программиста"

На основании частного, т.е. информации о некоторых свойствах класса "возраст < 20 лет" и "желаемый уровень вознаграждения > 700 условных единиц", мы делаем вывод об общем, а именно: соискатели - программисты.

Прогностическое моделирование, напротив, дедуктивно. Закономерности, полученные на этой стадии, формируются от общего к частному и единичному. Здесь мы получаем новое знание о некотором объекте или же группе объектов на основании:

- знания класса, к которому принадлежат исследуемые объекты;
- знание общего правила, действующего в пределах данного класса объектов.

Знаем, что соискатель ищет руководящую работу и его стаж > 15 лет, на 65% можно быть уверенным в том, что возраст соискателя > 35 лет.

На основании некоторых общих правил, а именно: цель соискателя - руководящая работа и его стаж > 15 лет, мы делаем вывод о единичном - возраст соискателя > 35 лет.

Следует отметить, что полученные *закономерности,* а точнее, их конструкции, могут быть прозрачными, т.е. допускающими толкование аналитика (рассмотренные выше правила), и непрозрачными, так называемыми "черными ящиками". Типичный пример последней конструкции - нейронная сеть.

3. ***Анализ исключений*** (forensic analysis)

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

Действие, выполняемое на этой стадии, - выявление отклонений (*deviation* detection). Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии *свободного поиска*.

Вернемся к одному из примеров, рассмотренному выше.

Найдено правило "Если возраст > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90 % случаев соискатель ищет руководящую работу". Возникает вопрос - к чему отнести оставшиеся 10 % случаев?

Здесь возможно два варианта. Первый из них - существует некоторое логическое объяснение, которое также может быть оформлено в виде правила. Второй вариант для оставшихся 10% - это ошибки исходных данных. В этом

случае стадия *анализа исключений* может быть использована в качестве *очистки данных*.

Классификация методов Data Mining

Далее мы рассмотрим несколько известных классификаций *методов Data Mining* по различным признакам.

Классификация технологических методов Data Mining

Все *методы Data Mining* подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после *Data Mining* либо они дистиллируются для последующего использования.

1. Непосредственное использование данных, или *сохранение данных*.

В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях *прогностического моделирования* и/или *анализа исключений*. Проблема этой группы *методов* - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. Выявление и использование формализованных *закономерностей*, или *дистилляция шаблонов*.

При технологии ***дистилляции шаблонов*** один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого *метода Data Mining*. Этот процесс выполняется на стадии *свободного поиска*, у первой же группы *методов* данная стадия в принципе отсутствует. На стадиях *прогностического моделирования* и *анализа исключений* используются результаты стадии *свободного поиска*, они значительно компактнее самих баз данных. Напомним, что конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

Методы этой группы: логические *методы* ; *методы* визуализации; *методы* кросс-табуляции; *методы*, основанные на уравнениях.

Логические *методы*, или *методы* логической индукции, включают: нечеткие запросы и анализы; символьные правила; деревья решений; генетические *алгоритмы*.

Методы этой группы являются, пожалуй, наиболее интерпретируемыми - они оформляют найденные *закономерности*, в большинстве случаев, в достаточно прозрачном виде с точки зрения пользователя. Полученные правила могут включать непрерывные и дискретные переменные. Следует заметить, что деревья решений могут быть легко преобразованы в наборы символьных правил путем генерации одного правила по пути от корня дерева до его *терминальной вершины*. Деревья решений и правила фактически являются разными способами решения одной задачи и отличаются лишь по своим возможностям. Кроме того, реализация правил осуществляется более медленными *алгоритмами*, чем индукция деревьев решений.

Методы кросс-табуляции: агенты, баесовские (доверительные) сети, кросс-табличная визуализация. Последний метод не совсем отвечает одному из свойств *Data Mining* - самостоятельному поиску *закономерностей* аналитической системой. Однако, предоставление информации в виде кросс-таблиц обеспечивает реализацию основной задачи *Data Mining* - поиск шаблонов, поэтому этот *метод* можно также считать одним из *методов Data Mining*.

Методы на основе уравнений.

Методы этой группы выражают выявленные *закономерности* в виде математических выражений - уравнений. Следовательно, они могут работать лишь с численными переменными, и переменные других типов должны быть закодированы соответствующим образом. Это несколько ограничивает применение *методов* данной группы, тем не менее они широко используются при решении различных задач, особенно задач прогнозирования.

Основные *методы* данной группы: статистические *методы* и нейронные сети

Статистические *методы* наиболее часто применяются для решения задач прогнозирования. Существует множество *методов* статистического анализа данных, среди них, например, корреляционно-регрессионный анализ, корреляция рядов динамики, выявление тенденций динамических рядов, гармонический анализ.

Другая классификация разделяет все многообразие *методов Data Mining* на две группы: статистические и кибернетические *методы*. Эта схема разделения основана на различных подходах к обучению математических моделей.

Следует отметить, что существует два подхода отнесения статистических *методов* к *Data Mining*. Первый из них противопоставляет статистические *методы* и *Data Mining*, его сторонники считают классические статистические *методы* отдельным направлением анализа данных. Согласно второму подходу,

статистические *методы* анализа являются частью математического инструментария *Data Mining*. Большинство авторитетных источников придерживается второго подхода.

В этой классификации различают две группы *методов*:

- статистические *методы*, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
- кибернетические *методы*, включающие множество разнородных математических подходов.

Недостаток такой классификации: и статистические, и кибернетические *алгоритмы* тем или иным образом опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к *извлечению знаний* из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах *Data Mining*.

Рассмотрим подробнее представленные выше группы.

Статистические методы Data mining

В эти *методы* представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- выявление связей и *закономерностей* (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, *факторный анализ* и др.);
- *динамические модели* и прогноз на основе временных рядов.

Арсенал статистических *методов Data Mining* классифицирован на четыре группы *методов*:

1. Дескриптивный анализ и описание исходных данных.
2. *Анализ связей* (корреляционный и регрессионный анализ, *факторный анализ*, *дисперсионный анализ*).

3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).

4. Анализ временных рядов (*динамические модели* и прогнозирование).

Кибернетические методы Data Mining

Второе направление *Data Mining* - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие *методы*:

- *искусственные нейронные сети* (распознавание, кластеризация, прогноз);
- *эволюционное программирование* (в т.ч. *алгоритмы* метода группового учета аргументов);
- генетические *алгоритмы* (оптимизация);
- *ассоциативная память* (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

Методы *Data Mining* также можно классифицировать по задачам *Data Mining*.

В соответствии с такой классификацией выделяем две группы. Первая из них - это подразделение *методов Data Mining* на решающие задачи сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования.

В соответствии со второй классификацией по задачам *методы Data Mining* могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные *методы* служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные *методы* кластерного анализа, в том числе: *алгоритм* k-средних, k-медианы, иерархические *методы* кластерного анализа, *самоорганизующиеся карты* Кохонена, *методы* кросс-табличной визуализации, различные *методы* визуализации и другие.

Прогнозирующие *методы* используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие *методы*: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод *опорных векторов* и др.

Свойства методов Data Mining

Различные *методы Data Mining* характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств.

Среди основных свойств и характеристик *методов Data Mining* рассмотрим следующие: точность, *масштабируемость*, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

Масштабируемость - свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и пр., при добавлении к ней вычислительных ресурсов.

В таблице 3.1 приведена сравнительная характеристика некоторых распространенных *методов*. Оценка каждой из характеристик проведена следующими категориями, в порядке возрастания: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/высокая, высокая, очень высокая.

Таблица 3.1. Сравнительная характеристика методов *Data Mining*

Алгоритм	Точность	Масштабируемость	Интерпретируемость	Пригодность к использованию	Трудоемкость	Разносторонность	Быстрота	Популярность, широта использования
классические *методы* (линейная регрессия)	нейтральная	высокая	высокая / нейтральная	высокая	нейтральная	нейтральная	высокая	низкая
нейронные сети	высокая	низкая	низкая	низкая	нейтральная	низкая	очень низкая	низкая
методы визуализации	высокая	очень низкая	высокая	высокая	очень высокая	низкая	чрезвычайно низкая	высокая / нейтральная
деревья решений	низкая	высокая	высокая	высокая / нейтральная	высокая	высокая	высокая / нейтральная	высокая / нейтральная
полиномиальные нейронные сети	высокая	нейтральная	низкая	высокая / нейтральная	нейтральная / низкая	нейтральная	низкая / нейтральная	нейтральная

k-ближайшего соседа	низкая	очень низкая	высокая / нейтральная	нейтральная	нейтральная / низкая	низкая	высокая	низкая

Как видно из рассмотренной таблицы, каждый из *методов* имеет свои сильные и слабые стороны. Но ни один метод, какой бы не была его оценка с точки зрения присущих ему характеристик, не может обеспечить решение всего спектра задач *Data Mining*.

Большинство инструментов *Data Mining*, предлагаемых сейчас на рынке программного обеспечения, 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 (например, SPSS, SAS, STATGRAPHICS, Statistica, др.) 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.