

Министерство образования и науки Украины
Донбасская государственная машиностроительная академия

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

КОНСПЕКТ ЛЕКЦИЙ

(для студентов специальности
«Системы и методы принятия решений»)

Утверждено
на заседании кафедры ИСПР
Протокол № 2 от 9 сентября 2014г.

Краматорск 2014

УДК 681.3

Интеллектуальный анализ данных : конспект лекций (для студентов специальности «Системы и методы принятия решений» всех форм обучения) / Сост. А.Ю. Мельников – Краматорск : ДГМА, 2014. – 44 с.

Приведены лекционные материалы к курсу.

Составитель	Мельников А.Ю., канд. техн. наук, доцент
-------------	--

Отв. за выпуск	Мельников А.Ю., канд. техн. наук, доцент
----------------	--

СОДЕРЖАНИЕ

Лекция №1. Введение в Data Mining.....	4
Лекция №2. Задачи и области применения Data Mining.....	9
Лекция №3. Методы классификации и прогнозирования.....	17
Лекция №4. Методы поиска ассоциативных правил и методы визуализации.....	30
Лекция №5. Этапы процесса Data Mining.....	35
Лекция №6. Организационные и человеческие факторы Data Mining...	41
Лекция №7. Стандарты и инструменты Data Mining.....	42

Модуль 1

Задачи и методы интеллектуального анализа данных

Лекция 1. Введение в Data Mining

Основные определения

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомым ценностей.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний в базах данных, информационная проходка данных, "промывание" данных. Понятие "обнаружение знаний в базах данных" (Knowledge Discovery in Databases, KDD) можно считать синонимом Data Mining. Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов.

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации).

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.

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных - это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

Знания - совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д.

Data Mining - это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.

Data Mining - это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур

(patterns) с целью достижения преимуществ в бизнесе (определение SAS Institute).

Data Mining - это процесс, цель которого - обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образов плюс применение статистических и математических методов (определение Gartner Group).

В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные подвыборкам данных, кои могут быть выражены в форме, понятной человеку.

Ограничения присущие Data Mining

1. Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

2. Сложность разработки и эксплуатации приложения Data Mining. Поскольку данная технология является мультидисциплинарной областью, для разработки приложения, включающего Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечить их качественное взаимодействие.

3. Квалификация пользователя. Различные инструменты Data Mining имеют различную степень "дружелюбности" интерфейса и требуют определенной квалификации пользователя. Поэтому программное обеспечение должно соответствовать уровню подготовки пользователя.

4. Извлечение полезных сведений невозможно без хорошего понимания сути данных. Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining.

5. Сложность подготовки данных. Успешный анализ требует качественной предобработки данных.

6. Большой процент ложных, недостоверных или бессмысленных результатов.

7. Высокая стоимость.

8. Наличие достаточного количества репрезентативных данных.

Данные

Данные – это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Как правило, данные состоят из объектов. Объект описывается как набор атрибутов. Атрибут - свойство, характеризующее объект.

Переменная (variable) - свойство или характеристика, общая для всех изучаемых объектов, проявление которой может изменяться от объекта к объекту. Значение (value) переменной является проявлением признака.

При анализе данных, как правило, нет возможности рассмотреть всю интересующую нас совокупность объектов. Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть выборку, и получить интересующую нас информацию на ее основании. Однако размер выборки должен зависеть от разнообразия объектов, представленных в генеральной совокупности.

Генеральная совокупность (population) - вся совокупность изучаемых объектов, интересующая исследователя.

Выборка (sample) - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

Параметры - числовые характеристики генеральной совокупности.

Статистики - числовые характеристики выборки.

Часто исследования основываются на гипотезах. Гипотезы проверяются с помощью данных. Гипотеза - предположение относительно параметров совокупности объектов, которое должно быть проверено на ее части.

Гипотеза - частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

1 Номинальная шкала (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия. Пример такой шкалы: профессии, город проживания, семейное положение. Для этой шкалы применимы только такие операции: равно (=), не равно ().

2 Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними. Шкала измерений дает возможность ранжировать значения переменных. Пример такой шкалы: место (1, 2, 3-е), которое команда получила на соревнованиях, номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.). Для этой шкалы применимы только такие операции: равно (=), не равно (), больше (>), меньше (<).

3 Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла. Пример такой шкалы: температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше. Для этой шкалы применимы только такие операции: равно (=), не равно (), больше (>), меньше (<), операции сложения (+) и вычитания (-).

4 Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы. Пример такой шкалы: вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее. Для этой шкалы применимы только такие операции: равно (=), не равно (\neq), больше ($>$), меньше ($<$), операции сложения (+) и вычитания (-), умножения (*) и деления (/).

5 Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории. Пример такой шкалы: пол (мужской и женский).

Типы наборов данных

1. Данные, состоящие из записей
 - Табличные данные - данные, состоящие из записей, каждая из которых состоит из фиксированного набора атрибутов.
 - Транзакционные данные представляют собой особый тип данных, где каждая запись, являющаяся транзакцией, включает набор значений.
2. Графические данные
 - Графы
 - Карты
 - Гипертекстовые
3. Химические данные

Форматы хранения данных

Возможны четыре аспекта работы с данными: определение данных, вычисление, манипулирование и обработка (сбор, передача и др.).

При манипулировании данными используется структура данных типа "файл". Файлы могут иметь различные форматы.

Наиболее распространенные форматы, согласно опросу "Форматы хранения данных", представлены на рис. 1.

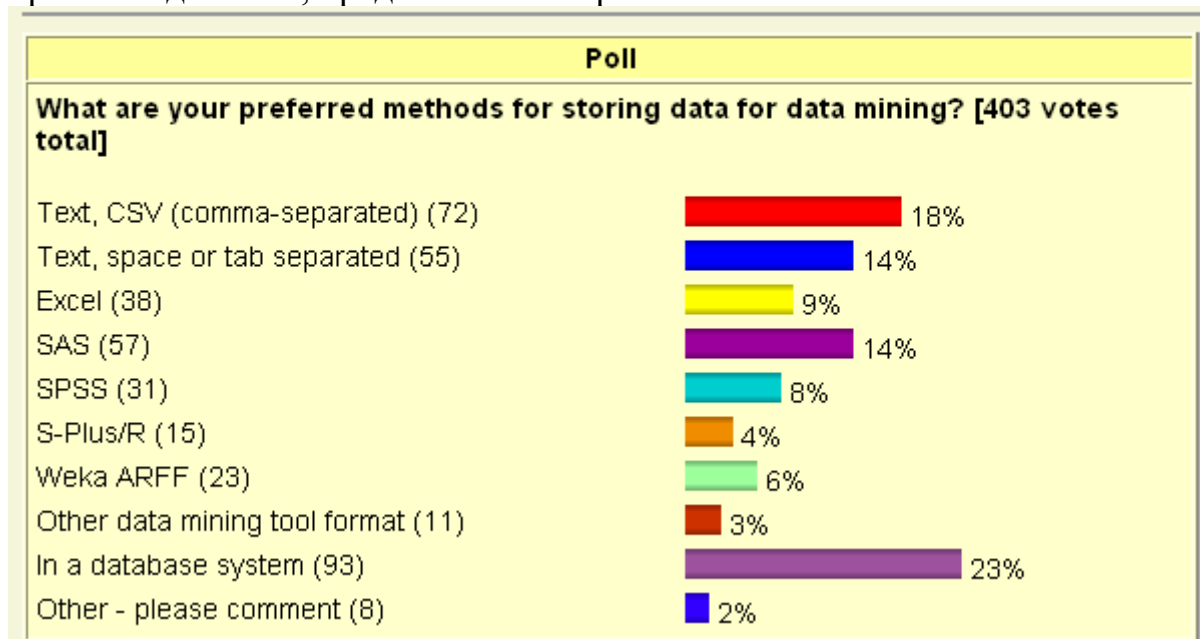


Рис. 1 – Наиболее распространенные форматы хранения данных

Методы и стадии Data Mining

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

Data Mining состоит из трех стадий.

Стадия 1. Выявление закономерностей (свободный поиск).

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

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

1. Свободный поиск (Discovery)

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей.

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

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).

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.

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.

Виды классификаций методов Data mining

1. по отношению к данным

- данные сохраняются после использования метода (метод кластерного анализа)
- данные корректируются для последующего использования (математические методы).

2. подходы к обучению модели

- статистические методы, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
 - кибернетические методы, включающие множество разнородных математических подходов.
3. по задачам Data mining
- задачи сегментации
 - задачи прогнозирования

Свойства методов Data mining: точность, масштабируемость, интерпретируемость, пригодность к использованию, трудоемкость, разносторонность, быстрота, популярность, широта использования.

Лекция №2. Задачи и области применения Data Mining

Задачи Data Mining

Классификация (Classification). 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. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).

Кластеризация (Clustering) является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не predetermined. Результатом кластеризации является разбиение объектов на группы. Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.

Ассоциация (Associations). В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Последовательность (Sequence), или последовательная ассоциация (sequential association). Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern).

Прогнозирование (Forecasting). В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для ре-

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

Определение отклонений или выбросов (Deviation Detection), анализ отклонений или выбросов. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Оценивание (Estimation). Задача оценивания сводится к предсказанию непрерывных значений признака.

Анализ связей (Link Analysis) - задача нахождения зависимостей в наборе данных.

Визуализация (Visualization, Graph Mining). В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

Подведение итогов (Summarization) - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

Виды классификаций задач

1 по стратегиям, задачи Data Mining подразделяются на следующие группы:

- обучение с учителем (классификация, оценка, прогнозирование);
- обучение без учителя (кластеризация);
- другие (входят задачи, не включенные в предыдущие две стратегии).

2 от типа используемых моделей

- дискриптивные (описательные). Аналитик получает шаблоны, описывающие данные, которые поддаются интерпретации;
- прогнозирующие (predictive) основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных.

3 по видам используемых моделей

- исследование и открытие
- прогнозирование и классификация
- объяснение и описание

Связь понятий

Рассмотрим два потока:

ДАННЫЕ - ИНФОРМАЦИЯ - ЗНАНИЯ И РЕШЕНИЯ

ЗАДАЧИ - ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ - ПРИЛОЖЕНИЯ

Эти потоки являются "двумя сторонами одной медали", отображением одного процесса, результатом которого должно быть знание и принятие решения.

Для начала рассмотрим первый поток. На рис. 2 показана связь понятий "данные", "информация" и "решения", которая возникает в процессе принятия решений.

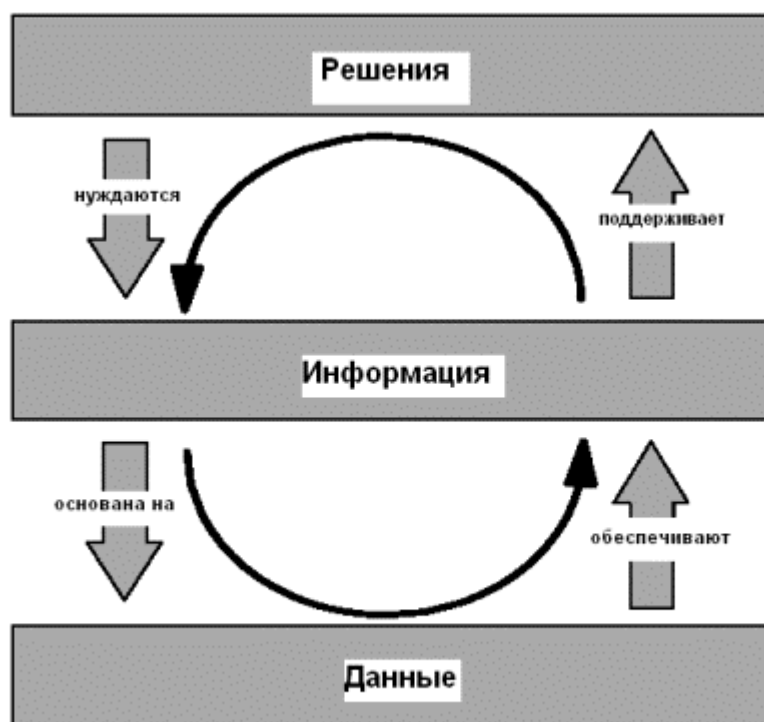


Рис 2 – Решения, информация и данные

Как видно из рисунка, данный процесс является циклическим. Принятие решений требует информации, которая основана на данных. Данные обеспечивают информацию, которая поддерживает решения, и т.д.

Рассмотренные понятия являются составной частью так называемой информационной пирамиды, в основании которой находятся данные, следующий уровень - это информация, затем идет решение, завершает пирамиду уровень знания. По мере продвижения вверх по информационной пирамиде объемы данных переходят в ценность решений, т.е. ценность для бизнеса. А, как известно, целью Business Intelligence является преобразование объемов данных в ценность бизнеса.

Следует отметить, что уровни анализа (данные, информация, знания) практически соответствуют этапам эволюции анализа данных, которая происходила на протяжении последних лет.

Верхний - уровень приложений - является уровнем решения.

Средний - уровень действий - по своей сути является уровнем информации, именно на нем выполняются действия Data Mining.

Нижний - уровень определения задачи Data Mining, которую необходимо решить применительно к данным, имеющимся в наличии.

Информация - любые, неизвестные ранее сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная интерпретация.

Свойства информации

1. Полнота информации.

Это свойство характеризует качество информации и определяет достаточность данных для принятия решений, т.е. информация должна содержать весь необходимый набор данных.

2. Достоверность информации.

Информация может быть достоверной и недостоверной. В недостоверной информации присутствует информационный шум, и чем он выше, тем ниже достоверность информации.

3. Ценность информации.

Ценность информации не может быть абстрактной. Информация должна быть полезной и ценной для определенной категории пользователей.

4. Адекватность информации.

Это свойство характеризует степень соответствия информации реальному объективному состоянию. Адекватная информация - это полная и достоверная информация.

5. Актуальность информации.

Информация должна быть актуальной, т.е. не устаревшей. Это свойство информации характеризует степень соответствия информации настоящему моменту времени.

6. Ясность информации.

Информация должна быть понятна тому кругу лиц, для которого она предназначена.

7. Доступность информации.

Доступность характеризует меру возможности получить определенную информацию. На это свойство информации влияют одновременно доступность данных и доступность адекватных методов.

8. Субъективность информации.

Информация носит субъективный характер, она определяется степенью восприятия субъекта (получателя информации). Знания - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача.

Задачи классификации

Классификация - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация требует соблюдения следующих правил:

1. в каждом акте деления необходимо применять только одно основание;

2. деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;

3. члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;

4. деление должно быть последовательным.

Различают:

- вспомогательную (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- естественную классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть:

- простой - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");
- сложной - применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Классификатором называется некая сущность, определяющая, какому из предопределенных классов принадлежит объект по вектору признаков.

Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации.

Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое.

Для классификации используются различные методы. Основные из них:

1. классификация с помощью деревьев решений;
2. байесовская (наивная) классификация;
3. классификация при помощи искусственных нейронных сетей;
4. классификация методом опорных векторов;
5. статистические методы, в частности, линейная регрессия;
6. классификация при помощи метода ближайшего соседа;
7. классификация CBR-методом;
8. классификация при помощи генетических алгоритмов.

Оценка точности классификации может проводиться при помощи кросс-проверки. Кросс-проверка (Cross-validation) - это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством. Точность классификации тестового множества сравнивается с точностью классификации обучающего множества. Если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная модель прошла кросс-проверку.

Задача кластеризации

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не predetermined.

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

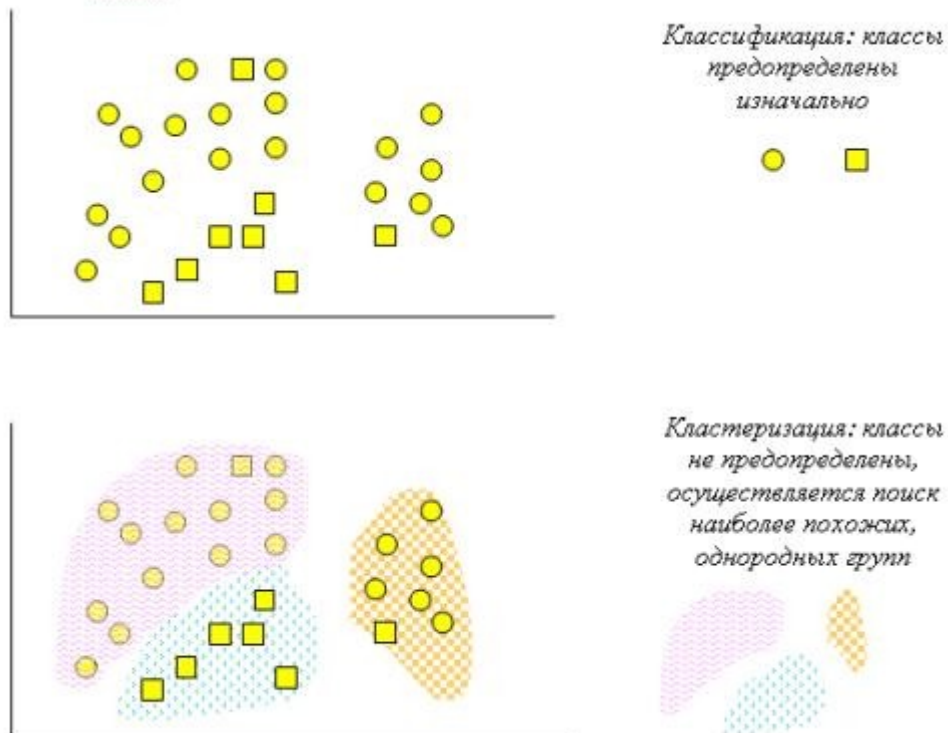
Цель кластеризации - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Кластер – группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.



Кластеры могут быть непересекающимися, или эксклюзивными (non-overlapping, exclusive), и пересекающимися (overlapping).

В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма.

Группы алгоритмов кластерного анализа:

- итеративные
- иерархические
- методы концентрации объектов

Задачи прогнозирования

Прогнозирование – это установление функциональной зависимости между зависимыми и независимыми переменными с целью определения неизвестных данных.

Временной ряд - последовательность наблюдаемых значений какого-либо признака, упорядоченных в неслучайные моменты времени.

В процессе определения структуры и закономерностей временного ряда предполагается обнаружение: шумов и выбросов, тренда, сезонной компоненты, циклической компоненты. Определение природы временного ряда может быть использовано как своеобразная "разведка" данных. Знание аналитика о наличии сезонной компоненты необходимо, например, для определения количества записей выборки, которое должно принимать участие в построении прогноза.

Тренд – это функция, которая формируется под действием тенденций влияющих на временной ряд.

Сезонность предполагает, что через определенные промежутки времени форма кривой, описывающей зависимую переменную, повторяет свои характерные очертания. Кроме сезонности ряд может быть циклическим. Циклы не имеют определенной продолжительности.

Период прогнозирования – основная единица времени, на которую делается прогноз.

Горизонт прогнозирования – число периодов в будущем, для которого делается прогноз.

Интервал прогнозирования – частота, с которой делается новый прогноз (может совпадать с периодом).

Виды прогнозов

Краткосрочный прогноз представляет собой прогноз на несколько шагов вперед, т.е. осуществляется построение прогноза не более чем на 3% от объема наблюдений или на 1-3 шага вперед.

Среднесрочный прогноз - это прогноз на 3-5% от объема наблюдений, но не более 7-12 шагов вперед; также под этим типом прогноза понимают прогноз на один или половину сезонного цикла.

Долгосрочный прогноз - это прогноз более чем на 5% от объема наблюдений.

Задача визуализации

Визуализация - это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения

Визуализации данных может быть представлена в виде: графиков, схем, гистограмм, диаграмм и т.д.

Сферы применения Data mining

Технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные.

Рассмотрим четыре основные сферы применения технологии Data Mining: наука, бизнес, исследования для правительства и Web-направление.

1 Применение Data Mining для решения бизнес-задач. Основные направления: банковское дело, финансы, страхование, CRM, производство, телекоммуникации, электронная коммерция, маркетинг, фондовый рынок и другие.

2 Применение Data Mining для решения задач государственного уровня. Основные направления: поиск лиц, уклоняющихся от налогов; средства в борьбе с терроризмом.

3 Применение Data Mining для научных исследований. Основные направления: медицина, биология, молекулярная генетика и геномная инженерия, биоинформатика, астрономия, прикладная химия, исследования, касающиеся наркотической зависимости, и другие.

4 Применение Data Mining для решения Web-задач. Основные направления: поисковые машины (search engines), счетчики и другие.

- Web Content Mining подразумевает автоматический поиск и извлечение качественной информации из разнообразных источников Интернета, перегруженных "информационным шумом". Здесь также идет речь о различных средствах кластеризации и аннотировании документов.
- Web Usage Mining подразумевает обнаружение закономерностей в действиях пользователя Web-узла или их группы.

Text Mining охватывает новые методы для выполнения семантического анализа текстов, информационного поиска и управления. Синонимом понятия Text Mining является KDT (Knowledge Discovering in Text - поиск или обнаружение знаний в тексте).

Технология Call Mining объединяет в себя распознавание речи, ее анализ и Data Mining. Ее цель - упрощение поиска в аудио-архивах, содержащих записи переговоров между операторами и клиентами. При помощи этой технологии операторы могут обнаруживать недостатки в системе обслуживания клиентов, находить возможности увеличения продаж, а также выявлять тенденции в обращениях клиентов.

Лекция №3. Методы классификации и прогнозирования

Основы анализа данных

К статистическому анализу относятся описательная статистика, корреляционный и регрессионный анализ.

Описательная статистика (Descriptive statistics) - техника сбора и суммирования количественных данных, которая используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения.

Цель описательной статистики - обобщить первичные результаты, полученные в результате наблюдений и экспериментов.

В состав описательной статистики входят такие характеристики: среднее; стандартная ошибка; медиана; мода; стандартное отклонение; дисперсия выборки; эксцесс; асимметричность; интервал; минимум; максимум; сумма; счет.

Медиана - точная середина выборки, которая делит ее на две равные части по числу наблюдений.

Эксцесс показывает "остроту пика" распределения, характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение (пик заострен). Отрицательный эксцесс обозначает относительно сглаженное распределение (пик закруглен). Эксцесс нормального распределения равен нулю.

Асимметрия или асимметричность показывает отклонение распределения от симметричного. Если асимметрия существенно отличается от нуля, то распределение несимметрично, нормальное распределение абсолютно симметрично. Если распределение имеет длинный правый хвост, асимметрия положительна; если длинный левый хвост - отрицательна.

Выбросы (outliers) - данные, резко отличающиеся от основного числа данных.

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине. Коэффициент корреляции, всегда обозначаемый латинской буквой r , используется для определения наличия взаимосвязи между двумя свойствами.

Связь между признаками (по шкале Чеддока) может быть сильной, средней и слабой. Тесноту связи определяют по величине коэффициента корреляции, который может принимать значения от -1 до $+1$ включительно.

Регрессионный анализ позволяет получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Как правило, проводится в 3 этапа:

- 1 установление формы зависимости;
- 2 определение функции регрессии;

3 оценка неизвестных значений зависимой переменной:

- расчет пропущенных значений (интрополяция);
- расчет будущих значений (экстраполяция).

Величина R-квадрат, называемая также мерой определенности, характеризует качество регрессионного анализа. В простом линейном регрессионном анализе множественный R равен коэффициенту корреляции Пирсона.

Методы классификации и прогнозирования. Деревья решений

Метод деревьев решений (decision trees) является методом решения задач классификации и прогнозирования.

Если зависимая, т.е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации. Если зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает задачу численного прогнозирования.

Внутренний узел дерева является узлом проверки, т.е. условием. При положительном ответе на вопрос осуществляется переход к левой части дерева, называемой левой ветвью, при отрицательном - к правой части дерева. Таким образом, внутренний узел дерева является узлом проверки определенного условия. Конечный узел дерева является узлом решения.

Конечный узел так же называют листом или вершиной. Если решается задача бинарной классификации, то модель называется дихотомической. Ветвление может быть в двух направлениях – да, нет.

Преимущества деревьев решений

1 Интуитивность деревьев решений. Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи. Результат работы алгоритмов конструирования деревьев решений легко интерпретируется пользователем.

2 Деревья решений дают возможность извлекать правила из базы данных на естественном языке.

3 Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.

4 Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева.

5 Точность моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (статистические методы, нейронные сети).

6 Разработан ряд масштабируемых алгоритмов, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных.

7 Быстрый процесс обучения.

8 Большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки пропущенных значений.

9 Деревья решений работают и с числовыми, и с категориальными типами данных.

10 Деревья решений строят непараметрические модели. Таким образом, деревья решений способны решать такие задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

Процесс конструирования дерева решений состоит из двух этапов: создание и сокращение дерева.

Процесс создания дерева происходит сверху вниз, т.е. является нисходящим. В ходе процесса алгоритм должен найти такой критерий расщепления, иногда также называемый критерием разбиения, чтобы разбить множество на подмножества, которые бы ассоциировались с данным узлом проверки. Каждый узел проверки должен быть помечен определенным атрибутом. Существует правило выбора атрибута: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению. Последняя фраза означает, что количество объектов из других классов, так называемых "примесей", в каждом классе должно стремиться к минимуму.

Существуют различные критерии расщепления. Наиболее известные - мера энтропии и индекс Gini.

Остановка - такой момент в процессе построения дерева, когда следует прекратить дальнейшие ветвления.

Первое правило остановки - "ранняя остановка" (prepruning), она определяет целесообразность разбиения узла.

Второе правило остановки обучения - ограничение глубины дерева. В этом случае построение заканчивается, если достигнута заданная глубина.

Третье правило остановки - задание минимального количества примеров, которые будут содержаться в конечных узлах дерева. При этом варианте ветвления продолжаются до того момента, пока все конечные узлы дерева не будут чистыми или будут содержать не более чем заданное число объектов.

Решением проблемы слишком ветвистого дерева является его сокращение путем отсечения (pruning) некоторых ветвей.

Качество классификационной модели, построенной при помощи дерева решений, характеризуется двумя основными признаками: точностью распознавания и ошибкой.

Точность распознавания рассчитывается как отношение объектов, правильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Ошибка рассчитывается как отношение объектов, неправильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Отсечение ветвей или замену некоторых ветвей поддеревом следует проводить там, где эта процедура не приводит к возрастанию ошибки. Процесс проходит снизу вверх, т.е. является восходящим.

На сегодняшний день существует большое число алгоритмов, реализующих деревья решений: CART, C4.5.

Алгоритм CART (Classification and Regression Tree) решает задачи классификации и регрессии. Атрибуты набора данных могут иметь как дискретное, так и числовое значение. Строит только бинарные деревья.

Алгоритм C4.5 строит дерево решений с неограниченным количеством ветвей у узла. Данный алгоритм может работать только с дискретным зависимым атрибутом и поэтому может решать только задачи классификации.

Метод опорных векторов

Метод опорных векторов (Support Vector Machine - SVM) относится к группе граничных методов. Она определяет классы при помощи границ областей. При помощи данного метода решаются задачи бинарной классификации. В основе метода лежит понятие плоскостей решений.

Плоскость (plane) решения разделяет объекты с разной классовой принадлежностью.

Цель метода опорных векторов - найти плоскость, разделяющую два множества объектов; такая плоскость показана на рис. 10.1 и 10.2. На этом рисунке множество образцов поделено на два класса: желтые объекты принадлежат классу A, коричневые - классу B.

Опорными векторами называются объекты множества, лежащие на границах областей.

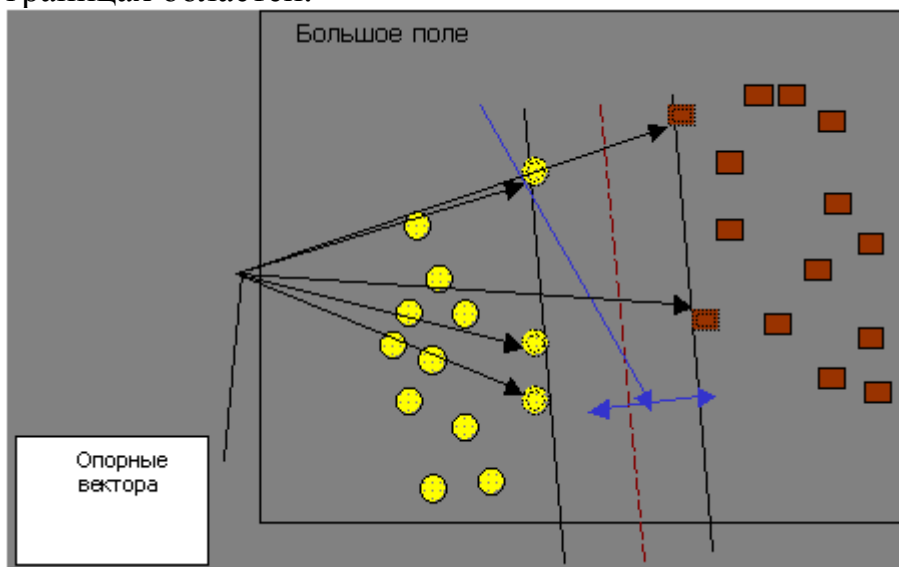


Рисунок 3 – Опорные вектора

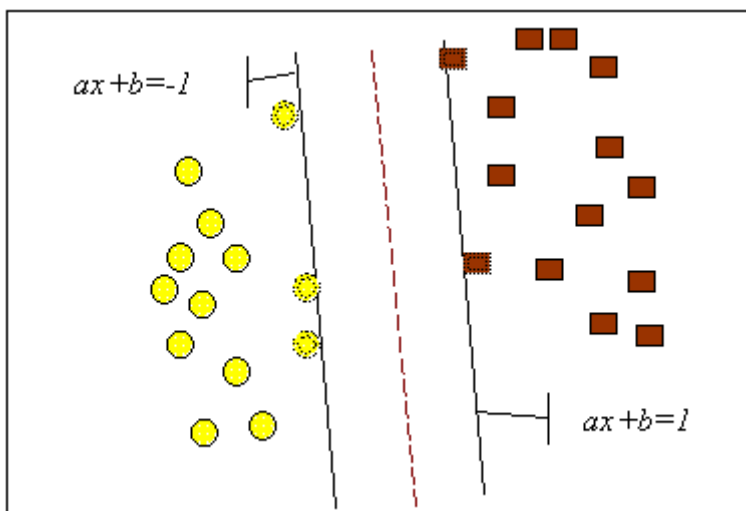


Рисунок 4 – Опорные вектора.

Задачу метода можно сформулировать как поиск функции $f(x)$, принимающей значения меньше нуля для векторов одного класса и больше нуля - для векторов другого класса.

Одной из проблем, связанных с решением задач классификации методом опорных векторов, является то обстоятельство, что не всегда можно легко найти линейную границу между двумя классами.

Недостаток метода состоит в том, что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Достоинство метода состоит в том, что для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных.

Метод ближайшего соседа

Метод "ближайшего соседа" ("nearest neighbour") основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.

При таком подходе используется термин "к-ближайший сосед" ("k-nearest neighbour"). Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей".

Данный метод по своей сути относится к категории "обучение без учителя". Качество зависит от наполненности базы данных.

Преимущества метода

- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки метода "ближайшего соседа"

- Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы.
- Существует сложность выбора меры "близости" (метрики). Также существует высокая зависимость результатов классификации от выбранной метрики.
- При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

С помощью данного метода решаются задачи классификации и регрессии.

Классификация новых объектов

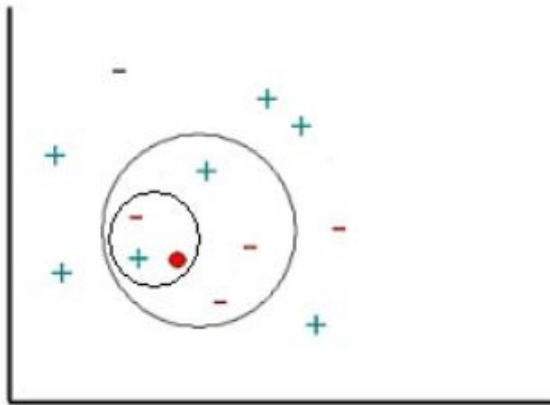


Рисунок 5 – Классификация объектов множества при разном значении параметра k

Результат работы метода с использованием одного ближайшего соседа. В этом случае отклик точки запроса будет классифицирован как знак плюс, так как ближайшая соседняя точка имеет знак плюс.

Увеличим число ближайших соседей до двух. Метод k -ближайших соседей не сможет классифицировать отклик точки запроса, поскольку вторая ближайшая точка имеет знак минус и оба знака равноценны.

Далее увеличим число используемых ближайших соседей до 5. Таким образом, будет определена целая окрестность точки запроса. Так как в области содержится 2 точки со знаком "+" и 3 точки со знаком "-", алгоритм k -ближайших соседей присвоит знак "-" отклику точки запроса.

В задаче прогнозирования решение получается усреднением k ближайших соседей:

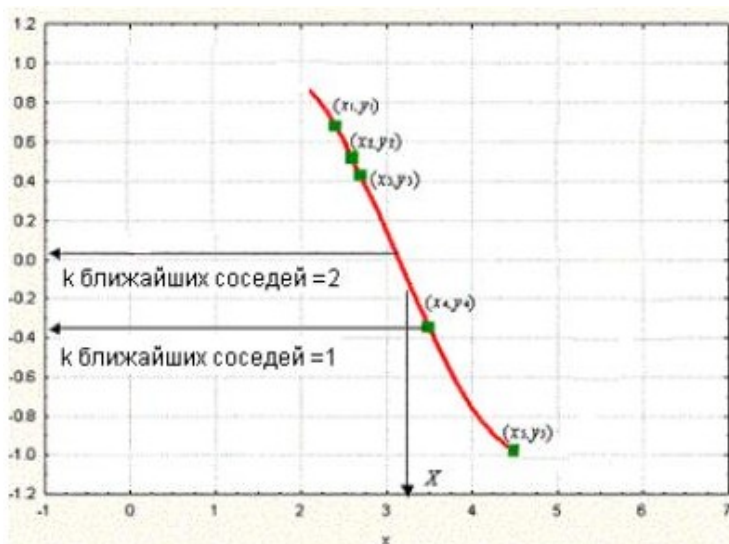


Рисунок 6 – Решение задачи прогнозирования при разных значениях параметра k

Байесовская классификация

Предполагает использование для решения задач классификации байесовских сетей. Исходит из предположения о взаимной независимости признаков, а так же об их одинаковой важности. Предполагается, что объект может принадлежать к определенному классу с заданной вероятностью. Основное применение метода Байесовской классификации – спам-фильтры.

Нейронные сети

Нейронные сети (Neural Networks) - это модели биологических нейронных сетей мозга, в которых нейроны имитируются относительно простыми, часто однотипными, элементами (искусственными нейронами).

Среди областей применения нейронных сетей - автоматизация процессов распознавания образов, прогнозирование, адаптивное управление, создание экспертных систем, организация ассоциативной памяти, обработка аналоговых и цифровых сигналов, синтез и идентификация электронных цепей и систем.

Модели нейронных сетей могут быть программного и аппаратного исполнения.

Среди задач Data Mining, решаемых с помощью нейронных сетей, будем рассматривать такие:

- Классификация (обучение с учителем). Примеры задач классификации: распознавание текста, распознавание речи, идентификация личности.
- Прогнозирование. Для нейронной сети задача прогнозирования может быть поставлена таким образом: найти наилучшее приближение функции, заданной конечным набором входных значений (обучающих примеров). Например, нейронные сети позволяют решать задачу восстановления пропущенных значений.

- Кластеризация (обучение без учителя). Примером задачи кластеризации может быть задача сжатия информации путем уменьшения размерности данных. Задачи кластеризации решаются, например, самоорганизующимися картами Кохонена. Этим сетям будет посвящена отдельная лекция.

Элементы нейронных сетей

Искусственный нейрон (формальный нейрон) - элемент искусственных нейронных сетей, моделирующий некоторые функции биологического нейрона.

Входные сигналы обрабатываются адаптивным сумматором, затем выходной сигнал сумматора поступает в нелинейный преобразователь, где преобразуется функцией активации, и результат подается на выход (в точку ветвления).

Аксон - выходную связь данного нейрона, с которой сигнал (возбуждения или торможения) поступает на синапсы следующих нейронов.

Активационная функция, которую также называют характеристической функцией, - это нелинейная функция, вычисляющая выходной сигнал формального нейрона.

Нелинейный преобразователь - это элемент искусственного нейрона, преобразующий текущее состояние нейрона (выходной сигнал адаптивного сумматора) в выходной сигнал нейрона по некоторому нелинейному закону (активационной функции).

Нейронные сети могут быть синхронные и асинхронные.

В синхронных нейронных сетях в каждый момент времени свое состояние меняет лишь один нейрон.

В асинхронных - состояние меняется сразу у целой группы нейронов, как правило, у всего слоя.

Однослойный персептрон (персептрон Розенблатта) - однослойная нейронная сеть, все нейроны которой имеют жесткую пороговую функцию активации.

Многослойный персептрон (MLP) - нейронная сеть прямого распространения сигнала (без обратных связей), в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько слоев.

Самоорганизующиеся карты Кохонена

Нейронные сети бывают с обратными связями и без обратных связей.

Сети с обратными связями

- Сети Хопфилда (задачи ассоциативной памяти).
- Сети Кохонена (задачи кластерного анализа).

Карты Кохонена (Self-Organizing Maps, SOM).

Основной принцип работы сетей Кохонена - введение в правило обучения нейрона информации относительно его расположения.

Самоорганизующиеся карты могут использоваться для решения таких задач, как моделирование, прогнозирование, поиск закономерностей в больших массивах данных, выявление наборов независимых признаков и сжатие информации.

Наиболее распространенное применение сетей Кохонена - решение задачи классификации без учителя, т.е. кластеризации.

Сеть Кохонена представляет собой два слоя: входной и выходной. Ее также называют самоорганизующей картой. Элементы карты располагаются в некотором пространстве, как правило, двумерном. Сеть Кохонена изображена на рис. 12.1.

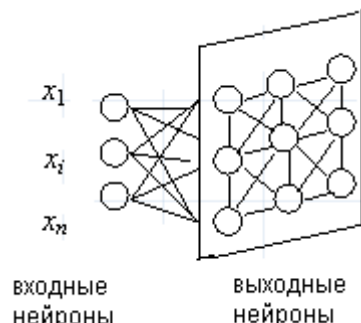


Рисунок 7 – Сеть Кохонена

Сеть Кохонена обучается методом последовательных приближений. В процессе обучения таких сетей на входы подаются данные, но сеть при этом подстраивается не под эталонное значение выхода, а под закономерности во входных данных. Начинается обучение с выбранного случайным образом выходного расположения центров.

Обучение при этом заключается не в минимизации ошибки, а в подстройке весов (внутренних параметров нейронной сети) для наибольшего совпадения с входными данными.

Темно-синие или светлые участки на карте соответствуют наименьшим значениям показателя, красные или темные - самым высоким.

Методы кластерного анализа. Иерархические методы

Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов, объединенных в две группы: иерархические и неиерархические методы.

Кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Как правило под расстоянием понимают евклидово расстояние между двумя точками i и j на плоскости, когда известны их координаты X и Y :

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

Кластер имеет следующие математические характеристики: центр, радиус, среднее квадратическое отклонение, размер кластера.

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от центра кластера.

Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют *спорными*.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по радиусу кластера, либо по среднее квадратическому отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков. В случае неоднородности единиц измерения признаков проводится предварительная стандартизация переменных.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES). В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA). Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рис. 8.

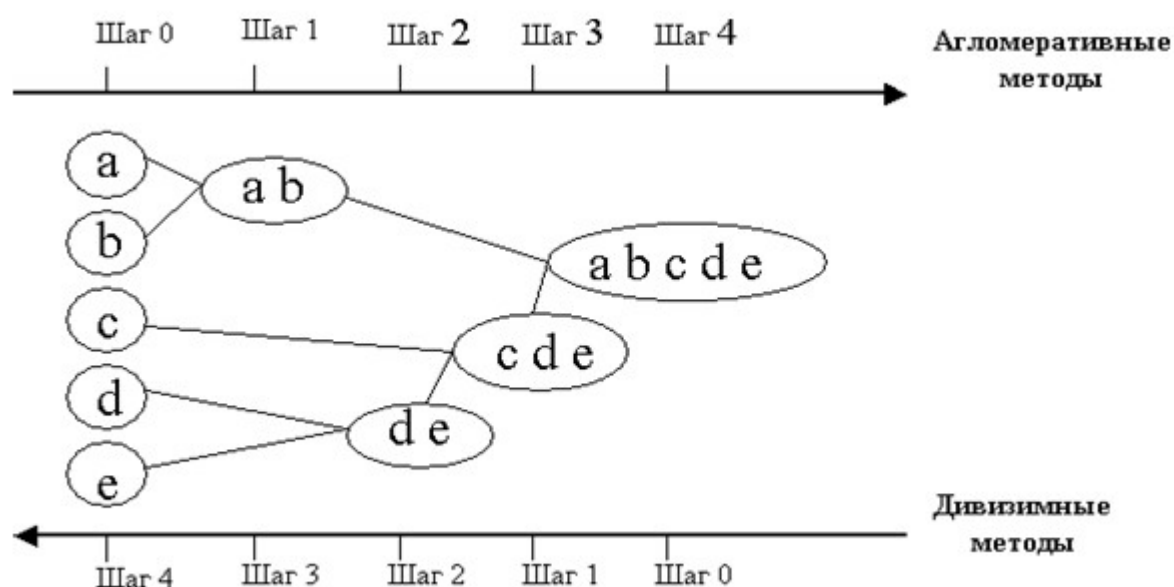


Рисунок 8 – Дендрограмма агломеративных и дивизимных методов

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия):

- 1 эвклидово расстояние;
- 2 квадрат эвклидова расстояния;
- 3 манхэттенское расстояние (расстояние городских кварталов). Это расстояние рассчитывается как среднее разностей по координатам;
- 4 расстояние Чебышева. Это расстояние используют, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению;
- 5 процент несогласия. Это расстояние вычисляется, если данные являются категориальными;

Методы объединения или связи

Метод ближнего соседа или **одиночная связь**. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.

Метод наиболее удаленных соседей или **полная связь**. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями").

Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до *центров кластеров*, получаемый в результате их объединения (Ward, 1963), используются методы дисперсионного анализа.

Метод невзвешенного попарного среднего. В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если кластеры неправильной формы

Метод взвешенного попарного среднего. В качестве весового коэффициента используется размер кластера (число объектов, содержащихся в кластере).

Невзвешенный центроидный метод. В качестве расстояния между двумя кластерами в этом методе берется расстояние между их центрами тяжести.

Взвешенный центроидный метод. Для учета разницы между *размерами кластеров* (числе объектов в них), используются веса.

Итеративные методы

При большом количестве наблюдений используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности.

Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает *алгоритм k-средних*, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Описание алгоритма

1. Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начальных центроидов может осуществляться следующим образом:

- выбор k-наблюдений для максимизации начального расстояния;
- случайный выбор k-наблюдений;
- выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

2. Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются по координатным средним кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

Достоинства *алгоритма k-средних*:

- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

Недостатки *алгоритма k-средних*:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;
- алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

РАМ является модификацией *алгоритма k-средних*, алгоритмом k-медианы (k-medoids). Алгоритм менее чувствителен к шумам и выбросам данных, чем алгоритм k-means, поскольку медиана меньше подвержена влияниям выбросов.

Решением больших баз данных является:

- предварительное сокращение размерности, т. е. проводится факторный анализ для изучения взаимосвязи между переменными, после чего большое число переменных сводится к меньшему числу независимых факторов;
- использование не всей базы данных, а какой то ее части.

Существует ряд сложностей, которые следует продумать перед проведением кластеризации.

- Сложность выбора характеристик, на основе которых проводится кластеризация. Необдуманный выбор приводит к неадекватному разбиению на кластеры и, как следствие, - к неверному решению задачи.
- Сложность выбора метода кластеризации. Этот выбор требует неплохого знания методов и предпосылок их использования.
- Проблема выбора числа кластеров. Если нет никаких сведений относительно возможного числа кластеров, необходимо провести ряд экспериментов и, в результате перебора различного числа кластеров, выбрать оптимальное их число.
- Проблема интерпретации результатов кластеризации. Форма кластеров в большинстве случаев определяется выбором метода объединения.

Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). В этом алгоритме реализован двухэтапный процесс кластеризации. В ходе первого этапа формируется предварительный набор кластеров. На втором этапе к выявленным кластерам применяются другие алгоритмы кластеризации - пригодные для работы в оперативной памяти.

Алгоритм WaveCluster представляет собой алгоритм кластеризации на основе волновых преобразований. В начале работы алгоритма данные обобщаются путем наложения на пространство данных многомерной решетки. На дальнейших шагах алгоритма анализируются не отдельные точки, а обобщенные характеристики точек, попавших в одну ячейку решетки. В результате такого обобщения необходимая информация уместается в оперативной памяти. На последующих шагах для определения кластеров алгоритм применяет волновое преобразование к обобщенным данным.

Алгоритм CLARA (Clustering LARge Applications). Алгоритм CLARA был разработан Kaufmann и Rousseeuw в 1990 году для кластеризации данных в больших базах данных. Алгоритм CLARA извлекает множество образцов из базы данных. Кластеризация применяется к каждому из образцов, на выходе алгоритма предлагается лучшая кластеризация.

Алгоритм Clarans (Clustering Large Applications based upon RANdomized Search) формулирует задачу кластеризации как случайный поиск в графе. В результате работы этого алгоритма совокупность узлов графа представляет собой разбиение множества данных на число кластеров, определенное пользователем.

Лекция №4. Методы поиска ассоциативных правил и методы визуализации

Методы поиска ассоциативных правил

Целью поиска ассоциативных правил (association rule) является нахождение закономерностей между связанными событиями в базах данных.

Часто встречающиеся приложения с применением ассоциативных правил:

- розничная торговля: определение товаров, которые стоит продвигать совместно; выбор местоположения товара в магазине; анализ потребительской корзины; прогнозирование спроса;
- перекрестные продажи: если есть информация о том, что клиенты приобрели продукты А, Б и В, то какие из них вероятнее всего купят продукт Г?
- маркетинг: поиск рыночных сегментов, тенденций покупательского поведения;
- сегментация клиентов: выявление общих характеристик клиентов компании, выявление групп покупателей;
- оформление каталогов, анализ сбытовых кампаний фирмы, определение последовательностей покупок клиентов (какая покупка последует за покупкой товара А);
- анализ Web-логов.

Рыночная корзина – это набор товаров, приобретенных покупателем в рамках одной отдельно взятой транзакции.

Транзакция – это множество событий, которые произошли одновременно.

Транзакционная или операционная база данных (Transaction database) представляет собой двумерную таблицу, которая состоит из номера транзакции (TID) и перечня покупок, приобретенных во время этой транзакции.

TID - уникальный идентификатор, определяющий каждую сделку или транзакцию.

Таблица 1 – Транзакционная база данных

TID	Приобретенные покупки	→	TID	Приобретенные покупки
100	Хлеб, молоко, печенье		100	a, b, c
200	Молоко, сметана		200	b, d
300	Молоко, хлеб, сметана, печенье		300	b, a, d, c
400	Колбаса, сметана		400	e, d
500	Хлеб, молоко, печенье, сметана		500	a, b, c, d
600	Конфеты		600	f

Поддержкой называют количество или процент *транзакций*, содержащих определенный набор данных.

$SUP(abc)=3$.

Ассоциативное правило имеет вид: "Из события А следует событие В".

В результате такого вида анализа мы устанавливаем закономерность следующего вида: "Если в транзакции встретился набор товаров (или набор элементов) А, то можно сделать вывод, что в этой же *транзакции* должен появиться набор элементов В)" Установление таких закономерностей дает нам возможность находить очень простые и понятные правила, называемые *ассоциативными*.

Основными характеристиками *ассоциативного правила* являются *поддержка* и *достоверность* правила.

Правило имеет *поддержку* s , если $s\%$ *транзакций* из всего набора содержат одновременно наборы элементов А и В или, другими словами, содержат оба товара.

Достоверность правила показывает, какова вероятность того, что из события А следует событие В.

Если значение *поддержки* правила слишком велико, то в результате работы алгоритма будут найдены правила очевидные и хорошо известные. Слишком низкое значение *поддержки* приведет к нахождению очень большого количества правил, которые, возможно, будут в большей части необоснованными, но не известными и не очевидными для аналитика. Таким образом, необходимо определить такой интервал, "золотую середину", который с одной стороны обеспечит нахождение неочевидных правил, а с другой - их обоснованность.

Если уровень *достоверности* слишком мал, то ценность правила вызывает серьезные сомнения. Например, правило с *достоверностью* в 3% только условно можно назвать правилом.

Методы поиска ассоциативных правил

Алгоритм AIS. Первый алгоритм поиска *ассоциативных правил*, называвшийся AIS, (предложенный Agrawal, Imielinski and Swami) был в 1993 году. В алгоритме AIS кандидаты множества наборов генерируются и подсчитываются "на лету", во время сканирования базы данных.

Алгоритм SETM. Формирует кандидатов "на лету", основываясь на преобразованиях базы данных. Чтобы использовать стандартную операцию объединения языка SQL для формирования кандидата, SETM отделяет формирование кандидата от их подсчета.

Алгоритм Apriori (1999). Множество кандидатов в наборе данных сначала формируются из базы данных и только после этого формирования начинается подсчет. При дальнейшем подсчете кандидатов осуществляется отсеивание набора с уровнем меньше минимального. Отсечение происходит на основе предположения о том, что у часто встречающегося набора товара все подмножество должно быть часто встречающимся.

AprioriTid. Особенность этого алгоритма - то, что база данных D не используется для подсчета *поддержки* кандидатов набора товаров после первого прохода. С этой целью используется кодирование кандидатов, выполненное на предыдущих проходах. В последующих проходах размер закодированных наборов может быть намного меньше, чем база данных, и таким образом экономятся значительные ресурсы.

AprioriHybrid. Алгоритм AprioriHybrid объединяет лучшие свойства алгоритмов Apriori и AprioriTid. AprioriHybrid использует алгоритм Apriori в начальных проходах и переходит к алгоритму AprioriTid, когда ожидается, что закодированный набор первоначального множества в конце прохода будет соответствовать возможностям памяти.

Методы визуализации

В 1987 году были сформулированы соответствующие задачи направления визуализации.

К способам визуального или графического представления данных относятся графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т.д.

Традиционные методы визуализации могут находить следующее применение:

- представлять пользователю информацию в наглядном виде;
- компактно описывать закономерности, присущие исходному набору данных;
- снижать размерность или сжимать информацию;
- восстанавливать пробелы в наборе данных;
- находить шумы и выбросы в наборе данных.

Каждый из алгоритмов Data Mining использует определенный подход к визуализации. Все эти способы визуального представления или отображения данных могут выполнять одну из функций:

- являются иллюстрацией построения модели (например, представление структуры (графа) нейронной сети);
- помогают интерпретировать полученный результат;
- являются средством оценки качества построенной модели;
- сочетают перечисленные выше функции (дерево решений, дендрограмма).

Методы визуализации, в зависимости от количества используемых измерений, принято классифицировать на две группы:

- представление данных в одном, двух и трех измерениях;
- представление данных в четырех и более измерениях.

Представление данных в одном, двух и трех измерениях. К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением. Практически любой современный инструмент Data Mining включает способы визуального представления из этой группы.

В соответствии с количеством измерений представления это могут быть следующие способы:

- одномерное (univariate) измерение, или **1-D**;
- двумерное (bivariate) измерение, или **2-D**;
- трехмерное или проекционное (projection) измерение, или **3-D**.

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

Наиболее известные способы многомерного представления информации:

- *параллельные координаты*;
- *"лица Чернова"*;
- *лепестковые диаграммы*.

В *параллельных координатах* переменные кодируются по горизонтали, вертикальная линия определяет значение переменной. Пример набора данных, представленного в декартовых координатах и *параллельных координатах*, дан на рис. 16.1. Этот метод представления многомерных данных был изобретен Альфредом Инселбергом (Alfred Inselberg) в 1985 году.

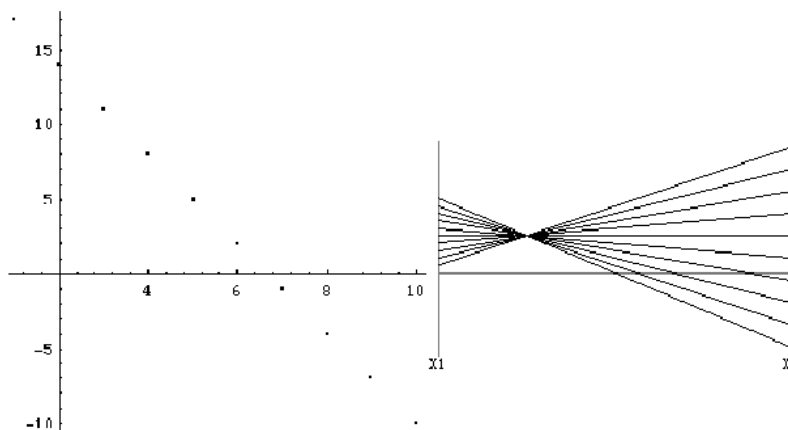


Рисунок 9 – Набор данных в декартовых и параллельных координатах

Лица Чернова. Основная идея представления информации состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица.

Принципы Тафта (Tufte's Principles) графического представления данных высокого качества гласят:

- предоставляйте пользователю самое большое количество идей, в самое короткое время, с наименьшим количеством чернил на наименьшем пространстве;
- говорите правду о данных.

Основные принципы компоновки визуальных средств представления информации:

1. Принцип лаконичности.
2. Принцип обобщения и унификации.
3. Принцип акцента на основных смысловых элементах.
4. Принцип автономности.
5. Принцип структурности.
6. Принцип стадийности.
7. Принцип использования привычных ассоциаций и стереотипов.

Модуль 2

Процессы и стандарты интеллектуального анализа данных

Лекция №5. Этапы процесса Data Mining

Комплексный подход к внедрению элементов СППР

Процесс Data Mining, который заключается в движении вверх по информационной пирамиде, неразрывно связан с процессом принятия решений, его можно рассматривать как неотъемлемую часть систем поддержки принятия решений (СППР).

Data Mining можно рассматривать как процесс поддержки принятия решений, при этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

СППР - интерактивная компьютерная система, предназначенная для поддержки принятия решений в слабоструктурированных и неструктурированных проблемах различных видов человеческой деятельности.

Выделяют следующие компоненты СППР:

- сервер хранилища данных;
- инструментарий OLAP;
- инструментарий Data Mining.

Эти компоненты СППР рассматривают такие основные вопросы: вопрос накопления данных и их моделирования на концептуальном уровне, вопрос эффективной загрузки данных из нескольких независимых источников и вопрос анализа данных.

Классификация СППР по сходству некоторых признаков (D.J. Power, 2000).

- СППР, ориентированные на данные (Data-driven DSS, Data-oriented DSS);
- СППР, ориентированные на модели (Model-driven DSS);
- СППР, ориентированные на знания (Knowledge-driven DSS);
- СППР, ориентированные на документы (Document-driven DSS);
- СППР, ориентированные на коммуникации и групповые СППР (Communications-Driven ? Group DSS);
- Интер-организованные и Интра-организованные СППР (Inter-Organizational или Intra-Organizational DSS);
- Специфически функциональные СППР или СППР общего назначения (Function-Specific или General Purpose DSS);
- СППР на базе Web (Web-Based DSS).

В зависимости от данных, с которыми работают СППР, выделяют два основных их типа СППР:

- EIS (Execution Information System) - информационная система Руководства, ИСР. СППР этого типа являются оперативными, ориентированы на неподготовленного пользователя, потому имеют упрощенный интерфейс,

базовый набор предлагаемых возможностей, фиксированные формы представления информации и перечень решаемых задач.

- DSS (Decision Support System). К системам этого типа относят многофункциональные системы анализа и исследования данных.

Общая схема поддержки принятия решений включает:

- помощь ЛПР при оценке состояния управляемой системы и воздействий на нее; выявление предпочтений ЛПР;
- генерацию возможных решений;
- оценку возможных альтернатив, исходя из предпочтений ЛПР;
- анализ последствий принимаемых решений и выбор лучшего с точки зрения ЛПР

OLAP-системы и OLAP-продукты

В основе концепции OLAP, или оперативной аналитической обработки данных (On-Line Analytical Processing), лежит многомерное концептуальное представление данных (Multidimensional conceptual view).

Термин OLAP введен Коддом (E. F. Codd) в 1993 году. Главная идея данной системы заключается в построении многомерных таблиц, которые могут быть доступны для запросов пользователей. Эти многомерные таблицы или так называемые многомерные кубы строятся на основе исходных и агрегированных данных.

Существует три способа хранения данных в OLAP-системах или три архитектуры OLAP-серверов:

- MOLAP (Multidimensional OLAP). Исходные и многомерные данные хранятся в многомерной БД или в многомерном локальном кубе. Такой способ хранения обеспечивает высокую скорость выполнения OLAP-операций. Но многомерная база в этом случае чаще всего будет избыточной. Куб, построенный на ее основе, будет сильно зависеть от числа измерений. При увеличении количества измерений объем куба будет экспоненциально расти.

- ROLAP (Relational OLAP). Исходные данные хранятся в реляционных БД или в плоских локальных таблицах на файл-сервере. Агрегатные данные могут помещаться в служебные таблицы в той же БД. Преобразование данных из реляционной БД в многомерные кубы происходит по запросу OLAP-средства. При этом скорость построения куба будет сильно зависеть от типа источника данных, и поэтому время отклика системы порой становится неприемлемо большим.

- HOLAP (Hybrid OLAP). Исходные данные остаются в реляционной базе, а агрегаты размещаются в многомерной. Построение OLAP-куба выполняется по запросу OLAP-средства на основе реляционных и многомерных данных. Такой подход позволяет избежать взрывного роста данных. При этом можно достичь оптимального времени исполнения клиентских запросов.

Следующая классификация - по месту размещения OLAP-машины. По этому признаку OLAP-продукты делятся на:

- OLAP-серверы (вычисления и хранение агрегатных данных выполняются отдельным процессом – сервером, клиентское приложение получает только результаты запросов к многомерным кубам, которые хранятся на сервере).

- OLAP-клиенты (построение многомерного куба и OLAP-вычисления выполняются в памяти клиентского компьютера).

Интеграция OLAP и Data Mining

Объединение понятий OLAP и Data mining может осуществляться тремя способами:

- "Cubing then mining". Возможность выполнения интеллектуального анализа должна обеспечиваться над любым результатом запроса к многомерному концептуальному представлению, то есть над любым фрагментом любой проекции гиперкуба показателей.

- "Mining then cubing". Подобно данным, извлеченным из хранилища, результаты интеллектуального анализа должны представляться в гиперкубической форме для последующего многомерного анализа.

- "Cubing while mining". Этот гибкий способ интеграции позволяет автоматически активизировать однотипные механизмы интеллектуальной обработки над результатом каждого шага многомерного анализа (перехода между уровнями обобщения, извлечения нового фрагмента гиперкуба и т.д.).

Хранилища данных

Если данные хранятся в базах данных различных информационных систем предприятия, при их анализе возникает ряд сложностей, в частности, значительно возрастает время, необходимое для обработки запросов; могут возникать проблемы с поддержкой различных форматов данных, а также с их кодированием; невозможность анализа длительных рядов ретроспективных данных и т.д.

Эта проблема решается путем создания хранилища данных. Задачей такого хранилища является интеграция, актуализация и согласование оперативных данных из разнородных источников для формирования единого непротиворечивого взгляда на объект управления в целом. На основе хранилищ данных возможно составление всевозможной отчетности, а также проведение оперативной аналитической обработки и Data Mining.

Билл Инмон (Bill Inmon) определяет хранилища данных как "предметно ориентированные, интегрированные, неизменяемые, поддерживающие хронологию наборы данных, организованные с целью поддержки управления" и призванные выступать в роли "единого и единственного источника истины", который обеспечивает менеджеров и аналитиков достоверной информацией, необходимой для оперативного анализа и принятия решений.

Предметная ориентация хранилища данных означает, что данные объединены в категории и сохраняются соответственно областям, которые они описывают, а не применениям, их использующим.

Интегрированность означает, что данные удовлетворяют требованиям всего предприятия, а не одной функции бизнеса. Этим хранилище данных

гарантирует, что одинаковые отчеты, сгенерированные для разных аналитиков, будут содержать одинаковые результаты.

Привязка ко времени означает, что хранилище можно рассматривать как совокупность "исторических" данных: возможно восстановление данных на любой момент времени. Атрибут времени явно присутствует в структурах хранилища данных.

Неизменность означает, что, попав один раз в хранилище, данные там сохраняются и не изменяются. Данные в хранилище могут лишь добавляться.

Цель Хранилищ Данных - обеспечить для организации "единый образ существующей реальности".

Хранилище данных представляет собой своеобразный накопитель информации о деятельности предприятия. Данные в хранилище представлены в виде многомерных структур под названием "звезда" или "снежинка".

Этапы процесса. Начальные этапы

Процесс Data Mining строит модель, а в процессе принятия решений эта модель эксплуатируется.

Рассмотрим традиционный процесс Data Mining. Он включает следующие этапы:

- анализ предметной области;
- постановка задачи;
- подготовка данных;
- построение моделей;
- проверка и оценка моделей;
- выбор модели;
- применение модели;
- коррекция и обновление модели.

Этап 1. Анализ предметной области

Предметная область - это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию.

В процессе изучения предметной области должна быть создана ее модель. Знания из различных источников должны быть формализованы при помощи каких-либо средств.

Этап 2. Постановка задачи

Постановка задачи Data Mining включает следующие шаги:

- формулировка задачи;
- формализация задачи.

Постановка задачи включает также описание статического и динамического поведения исследуемых объектов.

Этап 3. Подготовка данных

Подготовка данных является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов

всего процесса Data Mining. Кроме того на этап подготовки данных может быть потрачено до 80% всего времени, отведенного на проект.

Содержит следующие шаги

1. Определение и анализ требований к данным
2. Сбор данных (определяет необходимое количество данных)
3. Предварительная обработка данных. Предусматривает оценивание качества данных и при необходимости их очищение. Очистка предусматривает работу с пропущенными значениями (удаление объектов, расчет, игнорирование), обработка дублированных данных, обработка шумов и выбросов.

Очистка данных

Очистка данных занимается выявлением и удалением несоответствий данных с целью улучшения их качества. Как правило, состоит из 5 этапов:

- 1 анализ данных;
- 2 определение порядка и правил преобразования данных;
- 3 подтверждение;
- 4 преобразование;
- 5 противоток очищенных данных.

Все инструменты очистки данных принято объединять в 3 категории:

1. Средства анализа и модернизации данных. Обработывают данные с целью выявления ошибок, несоответствий и определения необходимых очищающих преобразований

2. Специальные средства очистки. Работают с конкретными областями - в основном это имена и адреса - или же с исключением дубликатов

3. Инструменты ETL. Предусматривает расширение, трансформацию и загрузку. Дают возможность пользователю определять функциональность API функций.

Инструменты очистки данных разделяются на две условные категории:

- универсальные системы, предназначенные для обслуживания всей базы данных целиком;
- верификаторы имени/адреса для очистки только данных о клиентах.

Определение качественной программы очистки данных состоит из четырех элементов. Программа должна:

- не затрагивать правильные данные;
- исправлять неверные;
- создавать небольшой по объему отчет о подозрительных записях;
- требовать минимальных затрат на установку, обслуживание и ручные проверки.

Инструменты очистки данных обычно выполняют одну или несколько из следующих функций:

- Парсинг.
- Стандартизация.

- Проверка допустимости.
- Улучшение.
- Согласование и консолидация.

Этапы работы с моделью

Моделирование - единственный к настоящему времени систематизированный способ увидеть варианты будущего и определить потенциальные последствия альтернативных решений, что позволяет их объективно сравнивать.

Модели бывают прогнозирующие (классификационные) и дескриптивные (описательные).

Примеры прогнозирующих моделей - это модели линейной регрессии (простейшие модели) и модели на основе нейронных сетей. С помощью дескриптивных моделей решают задачи поиска ассоциативных правил, задачи кластеризации, группировки, обобщения.

Этап 4. Построение модели

С использованием различных методов и алгоритмов строятся модели, описывающие процесс решения задачи. Выбираются методы или алгоритмы, которые приводят к приемлемым результатам. Data Mining является итеративным процессом. Итерация - это циклическая управляющая структура, она содержит выбор между альтернативами и следование избранной.

Этап 5. Проверка и оценка моделей

Проверка модели подразумевает проверку ее достоверности или адекватности. Эта проверка заключается в определении степени соответствия модели реальности. Оценка модели подразумевает проверку ее правильности. Оценка построенной модели осуществляется путем ее тестирования.

Этап 6. Выбор модели

Если в результате моделирования нами было построено несколько различных моделей, то на основании их оценки мы можем осуществить выбор лучшей из них. Основные характеристики модели, которые определяют ее выбор, - это точность модели и эффективность работы алгоритма.

Этап 7. Применение модели

После тестирования, оценки и выбора модели следует этап применения модели. На этом этапе выбранная модель используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining. Для классификационных и прогнозирующих моделей на этом этапе прогнозируется целевой (выходной) атрибут (target attribute).

Этап 8. Коррекция и обновление модели

Основными причинами являются следующие:

- изменились входящие данные или их поведение;
- появились дополнительные данные для обучения;
- изменились требования к форме и количеству выходных данных;
- изменились цели бизнеса, которые повлияли на критерии принятия решений;

- изменилось внешнее окружение или среда (макроэкономика, политическая ситуация, научно-технический прогресс, появление новых конкурентов и товаров и т.д.).

Погрешности в процессе Data Mining

- неверные или недостоверные исходные допущения
- ограниченные возможности при сборе необходимых данных
- неуверенность пользователей
- неоправданно высокая стоимость

Лекция №6. Организационные и человеческие факторы Data Mining

Организационные факторы

Необходимо создать определенную организационную окружающую среду.

- Поток данных (flow of Data) в организации должен быть приспособлен к Data Mining, т.е. сотрудники должны быть заинтересованы в открытом сотрудничестве по обмену информацией.
- Организационная культура подразумевает активное открытое сотрудничество по обмену информацией между отделами компании и ее сотрудниками.
- Деловая окружающая среда. Направлять Ваши действия по Data Mining должен бизнес. В организации должна присутствовать готовность открыть доступ к данным и показателям, а также к другим аспектам деятельности.

Человеческие факторы

Человеческий фактор при внедрении Data Mining - это наличие и квалификационное соответствие специалистов, готовых работать с Data Mining.

Специалисты компании, вовлеченные в процесс Data Mining, исполняют одну из ролей:

- Специалист предметной области (Domain experts) - специалист, имеющий знания о окружении бизнеса, процессах, заказчиках, клиентах, потребителях, конкурентах, т.е. о предметной области.
- Администратор баз данных (Database administrator) - специалист, имеющий знания о том, где и каким образом хранятся данные, как получить к ним доступ и как связать между собой эти данные. Администратор базы данных отвечает за выработку требований к базе данных, за ее проектирование, реализацию, эффективное использование и сопровождение.
- Специалист по добыче данных (Mining specialists) - специалист по анализу данных, который имеет, как минимум, основы статистических знаний.



Рисунок 10 – Роли в Data Mining

Часто в процесс также вовлечены другие специалисты по информационным технологиям и менеджеры проектов. Среди них могут быть:

- менеджер проектов (Project Manager);
- специалист по IT Архитектуре (IT Architect);
- специалист по Архитектуре Решений (Solution Architect);
- специалист по Архитектуре Данных (Data Architect);
- специалист по Моделированию данных (Data Modeler);
- эксперт Data Mining (Data Mining Expert);
- деловой Аналитик (Business Analyst).

Лекция №7. Стандарты и инструменты Data Mining

1. CRISP-DM.

The Cross Industrie Standard Process for Data Mining – Стандартный межотраслевой процесс Data Mining – является наиболее популярной и распространенной методологией. Членами консорциума CRISP-DM являются NCR, SPSS и DaimlerChrysler.

В соответствии со стандартом CRISP, Data Mining включает следующие фазы:

1. Осмысление бизнеса (Business understanding).
2. Осмысление данных (Data understanding).
3. Подготовка данных (Data preparation).
4. Моделирование (Modeling).
5. Оценка результатов (Evaluation).
6. Внедрение (Deployment).

К этому набору фаз иногда добавляют седьмой шаг - Контроль, он заканчивает круг.

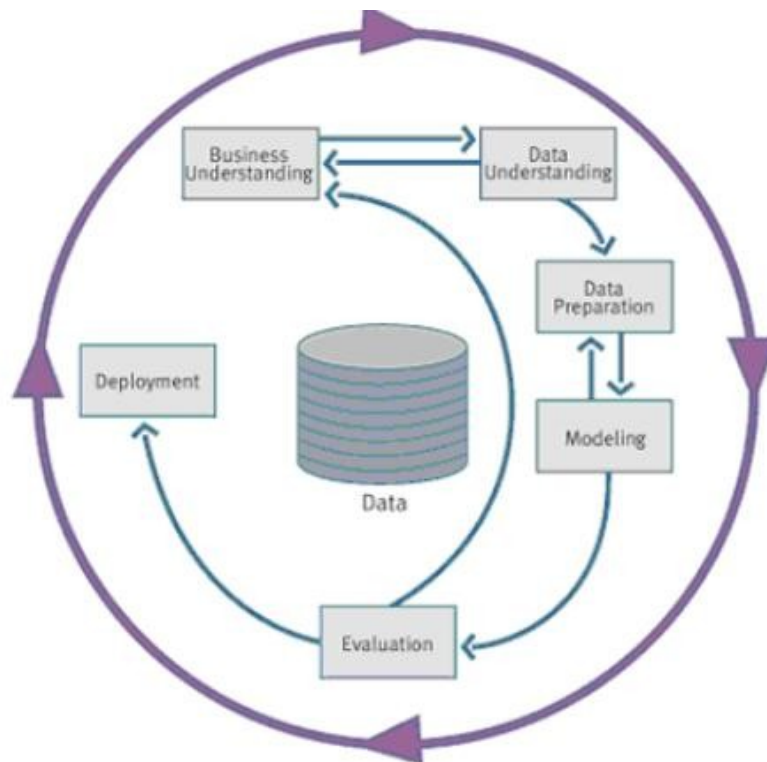


Рисунок 11 – Фазы, рекомендуемые моделью CRISP-DM

2. SEMMA методология реализована в среде SAS Data Mining Solution (SAS). Ее аббревиатура образована от слов Sample ("Отбор данных", т.е. создание выборки), Explore ("Исследование отношений в данных"), Modify ("Модификация данных"), Model ("Моделирование взаимосвязей"), Assess ("Оценка полученных моделей и результатов").

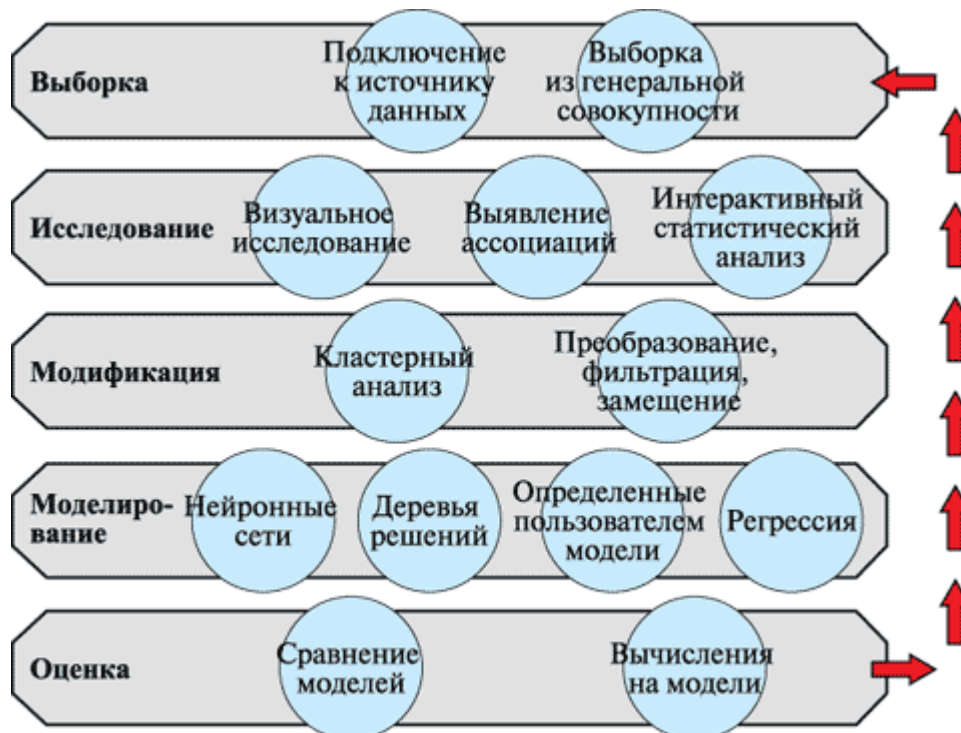


Рисунок 12 – Методология разработки проекта Data Mining в соответствии с методологией SEMMA

3. **PMML** (Predictive Modeling mark-up Language) – язык описания предикторных (или прогнозных) моделей или языке разметки для прогнозного моделирования. Основа этого стандарта - язык XML.

Основная цель стандарта PMML - обеспечение возможности обмена моделями данных между программным обеспечением разных разработчиков

При помощи стандарта PMML-совместимые приложения могут легко обмениваться моделями данных с другими PMML-инструментами. Таким образом, модель, созданная в одном программном продукте, может использоваться для прогнозного моделирования в другом.

Стандарт PMML включает:

- описание анализируемых данных (структура и типы данных);
- описание схемы анализа (используемые поля данных);
- описание трансформаций данных (например, преобразования типов данных);
- описание статистик, прогнозируемых полей и самих прогнозных моделей.

4. **JDM** (The Java Data Mining standard - Java Specification Request 73, JSR-73). Стандарт, разработанный группой JSR 73, Java Data Mining API (JDM) - это первая попытка создать стандартный Java API (программный интерфейс приложения) для получения доступа к инструментам Data Mining из Java-приложений.

5. **SQL/MM** представляет собой набор определенных пользователем SQL процедур для возможностей вычислений и использований моделей Data Mining

6. **The OLE DB for Data Mining standard of Microsoft**. Этот стандарт позволяет, подобно SQL/MM, применять методы Data Mining в структуре реляционных баз данных. Этот стандарт является расширением OLE DB

Стандарты, имеющие прямое или опосредованное отношение к Data Mining, можно объединить в группы:

- стандарты, базирующиеся на услугах Data Mining (услуги создания модели управления, скоринговые услуги, услуги анализа данных, услуги исследования данных, статистические услуги моделирования);
- стандарты web-службы (SOAP/XML, WSRF, и т.д), Grid-Услуги (OGSA, OGSA/DAI, и т.д.), Семантические Стандарты Web (RDF, OWL, и т.д.);
- стандарты, которые должны появиться в ближайшее время: стандарты для технологического процесса, стандарты для преобразований данных, стандарты для оперативного (real time) Data Mining, стандарты для сетей данных (data webs).