Data Mining

Интеллектуальный анализ данных

Добыча данных - Data Mining

Data Mining - исследование и обнаружение "машиной" (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

- Знания должны быть новые, ранее неизвестные.
- Знания должны быть нетривиальны.
- Знания должны быть практически полезны.
- Знания должны быть доступны для понимания человеку.

Задачи Data Mining

Задача классификации сводится к определению класса объекта по его характеристикам. Множество классов известно заранее.

Задача регрессии подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого параметра из множества действительных чисел.

При **поиске ассоциативных правил** целью является нахождение частых зависимостей (или ассоциаций)

Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных.

Описательные и предсказательные задачи

Описательные (descriptive) задачи уделяют внимание улучшению понимания анализируемых данных. К такому виду задач относятся кластеризация и поиск ассоциативных правил

Решение **предсказатедьных (predictive)** задач разбивается на два этапа. На первом этапе на основании набора данных с известными результатами строится модель. На втором этапе она используется для предсказания результатов на основании новых наборов данных. При этом, естественно, требуется, чтобы построенные модели работали максимально точно. К данному виду задач относят задачи классификации и регрессии. Сюда можно отнести и задачу поиска ассоциативных правил, если результаты ее решения могут быть использованы для предсказания появления некоторых событий.

Supervised и unsupervised learnig

В случае **supervised learning** задача анализа данных решается в несколько этапов. Сначала строится модель анализируемых данных - классификатор. Затем классификатор подвергается обучению. Другими словами, проверяется качество его работы, и, если оно неудовлетворительное, происходит дополнительное обучение классификатора. Так продолжается до тех пор, пока не будет достигнут требуемый уровень качества или не станет ясно, что выбранный алгоритм не работает корректно с данными, либо же сами данные не имеют структуры, которую можно выявить. К этому типу задач относят задачи классификации и регрессии.

Unsupervised learning объединяет задачи, выявляющие описательные модели. Например закономерности в покупках, совершаемых клиентами большого магазина. Достоинством таких задач является возможность их решения без каких либо предварительных знаний. об анализируемых данных. К этим задачам относятся кластеризация и поиск ассоциативных правил.

Задача классификации и регрессии

Требуется определить, к какому из известных классов относятся исследуемые объекты, т. е. классифицировать их.

Клиент банка: «кредитоспособен» и «некредитоспособен».

Фильтр электронной почты: «спам», «не спам»

Распознавание цифр: от 0 до 9.

В Data Mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров.

Задача классификации и регрессии решается в два этапа. На первом выделяется обучающая выборка. В нее входят объекты, для которых известны значения как независимых, так и зависимых переменных.

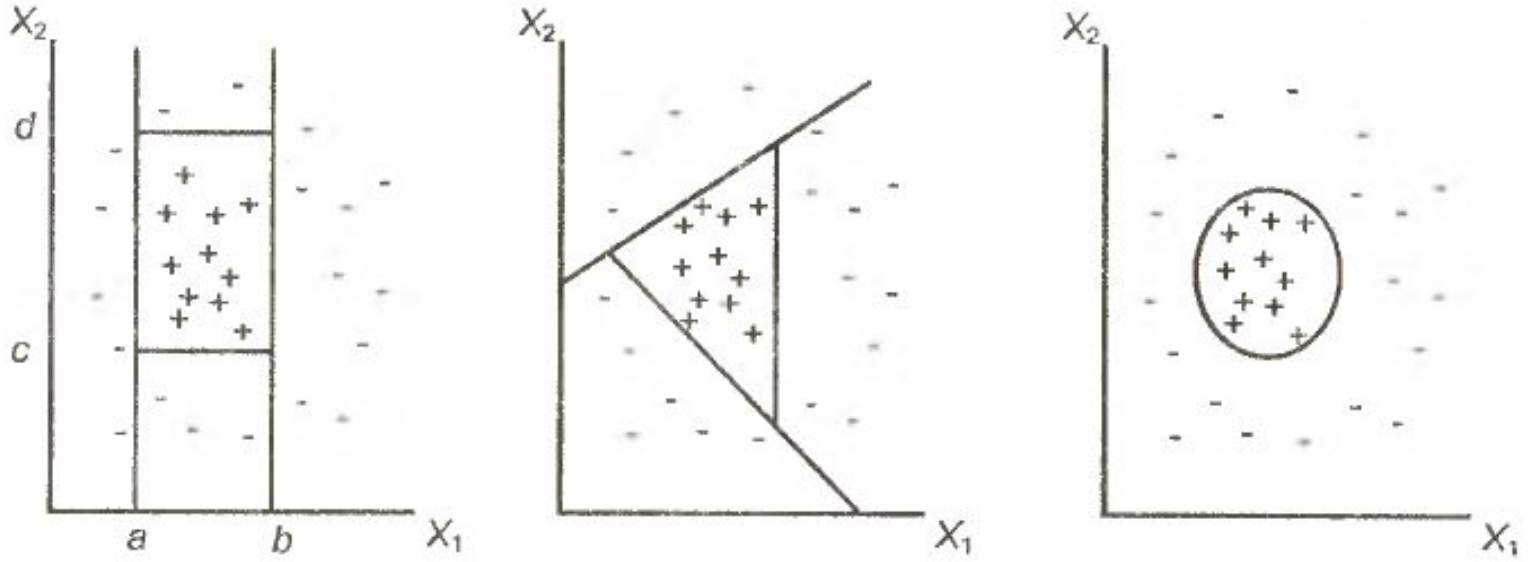
Задача классификации и регрессии

На основании обучающей выборки строится модель определения значения зависимой переменной. Ее часто называют функцией классификации или регрессии. Для получения максимально точной функции к обучающей выборке предъявляются следующие основные требования:

- количество объектов, входящих в выборку, должно быть достаточно большим. Чем больше объектов, тем точнее будет построенная на ее основе функция классификации или регрессии;
- в выборку должны входить объекты, представляющие все возможные классы в случае задачи классификации или всю область значений в случае задачи регрессии;
- для каждого класса в задаче классификации или для каждого интервала области значений в задаче регрессии выборка должна содержать достаточное количество объектов.

На втором этапе построенную модель применяют к анализируемым объектам (к объектам с неопределенным значением зависимой переменной).

Задача классификации и регрессии



Классификация в двумерном пространстве

Задача поиска ассоциативных правил

Суть задачи заключается в определении часто встречающихся наборов объектов в большом множестве таких наборов. Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах (анализ рыночных корзин - Basket Analysis). При анализе этих данных интерес прежде всего представляет информация о том, какие товары покупаются вместе, в какой последовательности, какие категории потребителей какие товары предпочитают, в какие периоды времени и т. п.

В сфере обслуживания интерес представляет информация о том, какими услугами клиенты предпочитают пользоваться в совокупности.

В медицине - анализ сочетания симптомов и болезней.

Сиквенциальный анализ учитывает последовательность происходящих событий (телекоммуникационные компании, анализ аварий).

Задача кластеризации

Задача кластеризации состоит в разделении исследуемого множества объектов на группы "похожих" объектов, называемых кластерами (cluster).

Периодическая система элементов Д.И. Менделеева.

Сегментация в маркетинге. Критериями сегментации являются: географическое местоположение, социально-демографические характеристики, мотивы совершения покупки и т. п.

На основании результатов сегментации маркетолог может определить, например, такие характеристики сегментов рынка, как реальная и потенциальная емкость сегмента, группы потребителей, чьи потребности не удовлетворяются в полной мере ни одним производителем, работающим на данном сегменте рынка, и т. п.

Практическое применение Data Mining

Интернет-технологии

- персонализация посетителей Web-сайтов
- поиск случаев мошенничества с кредитными картами
- Web Mining: Web content mining и Web usage mining

Торговля

- анализ рыночных корзин и сиквенциональный анализ

Телекоммуникации

- анализ доходности и риска потери клиентов
- защита от мошенничества,
- выявление категорий клиентов с похожими стереотипами пользования услугами и разработка привлекательных наборов цен и услуг

Практическое применение Data Mining

Промышленное производство

- прогнозирование качества изделия в зависимости от замеряемых параметров технологического процесса.

Медицина и биология

- построение диагностической системы
- исследование эффективности хирургического вмешательства
- Биоинформатика – изучение генов, разработка новых лекарств

Банковское дело

- оценка кредитоспособности заемщика

Модели Data Mining

Предсказательные модели

- модели классификации
- модели последовательностей

Описательные модели

- регрессионные модели
- модели кластеров
- модели исключений
- итоговые модели
- ассоциативные модели

Предсказательные модели

модели классификации описывают правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов. Такие правила строятся на основании информации о существующих объектах путем разбиения их на классы;

модели последовательностей описывают функции, позволяющие прогнозировать изменение непрерывных числовых параметров. Они строятся на основании данных об изменении некоторого параметра за прошедший период времени.

Описательные модели

регрессионные модели описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме. Они описывают функциональную зависимость не только между непрерывными числовыми параметрами, но и между категориальными параметрами;

модели кластеров описывают группы (кластеры), на которые можно разделить объекты, данные о которых подвергаются анализу. Группируются объекты (наблюдения, события) на основе данных (свойств), описывающих сущность объектов. Объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем сильнее "похожи" объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация;

Описательные модели

модели исключений описывают исключительные ситуации в записях (например, отдельных пациентов), которые резко отличаются чем либо от основного множества записей (группы больных). Знание исключений может быть использовано двояким образом. Возможно, эти записи представляют собой случайный сбой, например ошибки операторов, вводивших данные в компьютер. С другой стороны, отдельные исключительные записи могут представлять самостоятельный интерес для исследования, т. к. они могут указывать на некоторые редкие, но важные аномальные заболевания.

Описательные модели

итоговые модели - выявление ограничений на данные анализируемого массива. Например, при изучении выборки данных по пациентам не старше 30 лет, перенесшим инфаркт миокарда, обнаруживается, что все пациенты, описанные в этой выборке, либо курят более 5 пачек сигарет в день, либо имеют вес не ниже 95 Kr. Построение итоговых моделей заключается в нахождении каких либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей;

ассоциативные модели - выявление закономерностей между связанными событиями.

Методы Data Mining

Переборные алгоритмы, эвристики, статистические методы

Нечеткая логика

Генетические алгоритмы

Нейронные сети

Нечеткая логика

Неопределенность по объему отсутствующей информации у системного аналитика можно разделить на три большие группы:

1. Неизвестность.
2. Неполнота (недостаточность, неадекватность).
3. Недостоверность.

Недостоверность бывает **физической** (источником ее является внешняя среда)

и **лингвистической** (возникает в результате словесного обобщения и обусловливается необходимостью описания бесконечного числа ситуаций ограниченным числом слов в ограниченное время).

Неопределенность

Выделяют два вида **физической** неопределенности:

1. Неточность.
2. Случайность.

Для обработки физических неопределенностей успешно используются методы теории вероятностей и классическая теория множеств

Выделяют два вида **лингвистической** неопределенности:

1. Неопределенность значений слов (многозначность, расплывчатость, неясность, нечеткость).
2. Неоднозначность смысла фраз (выделяют синтаксическую и семантическую).

Нечеткая логика

Для работы с **лингвистической** неопределенности используют нечеткую логику (**теория нечетких множеств** - автор Лотфи Заде).

Заде предложил лингвистическую модель, которая использует не математические выражения, а слова, отражающие качество. Человеку в процессе управления сложными объектами свойственно оперировать понятиями и отношениями с расплывчатыми границами. Источником расплывчатости является существование классов объектов, степень принадлежности к которым величина, непрерывно изменяющаяся от полной принадлежности к нему до полной непринадлежности.

Основные особенности нечеткой логики:

1. Правила принятия решений являются условными высказываниями типа "если ... , то ... " и реализуются с помощью механизма логического вывода.
2. Вместо одного четкого обобщенного правила нечеткая логика оперирует со множеством частных правил.
3. Правила в виде "если ... , то ... " позволяют решать задачи классификации в режиме диалога с оператором, что способствует повышению качества классификатора уже в процессе эксплуатации.

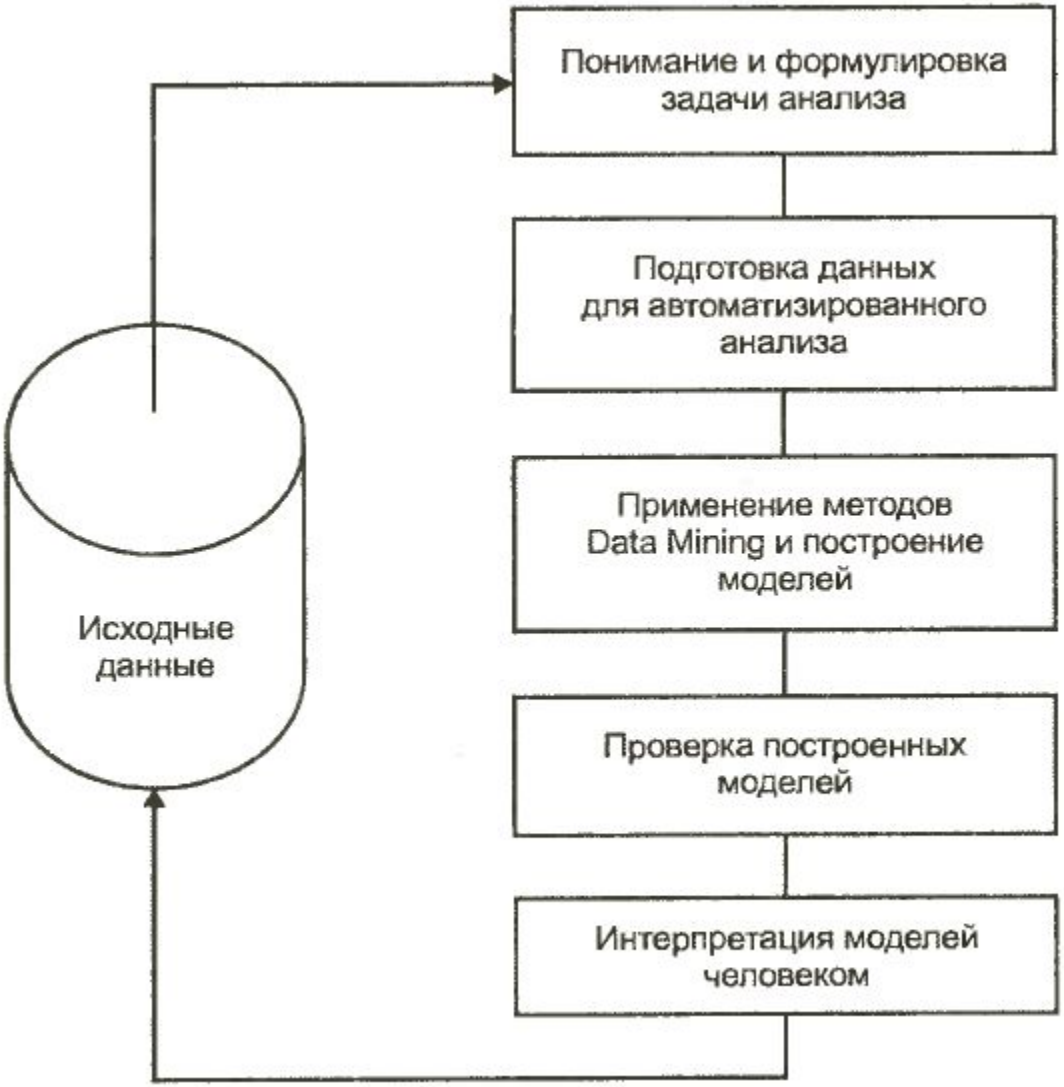
Генетические алгоритмы

Генети́ческий алгори́тм (англ. *genetic algorithm*) — это эвристический алгоритм) — это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую) — это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию) — это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию. Является разновидностью эволюционных вычислений) — это эвристический алгоритм поиска, используемый для решения задач

Нейронные сети

- **Иску́сственные нейро́нные се́ти** (ИНС) — математические модели (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы.
- ИНС представляют собой системуИНС представляют собой систему соединённых и взаимодействующих между собой простых процессоровИНС представляют собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейроновИНС представляют собой систему соединённых и

Процесс обнаружения знаний



Подготовка исходных данных

- выработать некий четкий набор числовых или нечисловых пара метров, характеризующих задачу,
- представить данные в виде таблицы,
- очистить данные по столбцам,
- очистить данные по строкам.

Средства Data Mining

- входящие, как неотъемлемая часть, в системы управления базами данных;
- библиотеки алгоритмов Data Mining с сопутствующей инфраструктурой;
- коробочные или настольные решения ("черные ящики").

Вопросы

- Что такое Data Mining?
- Основные задачи Data Mining.
- Описательные и предсказательные задачи.
- Supervised learning и unsupervised learning.
- Этапы интеллектуального анализа данных.
- Методы интеллектуального анализа данных.