Понятие интеллектуального анализа данных

Понятие интеллектуального анализа данных соответствует широко распространенному термину DataMining, который часто переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний в базах данных. Понятие DataMining, появившееся в 1978 году, приобрело большую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялся в рамках прикладной статистики, при этом в основном решались задачи по обработке небольших баз данных.

DataMining можно охарактеризовать как технологию, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей: - неочевидных, так как найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем; - объективных, так как обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным; - практически полезных, так как выводы имеют конкретное значение, которому можно найти практическое применение.

Набор данных и их атрибутов

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты. Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций. Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки. Иными словами, данные – это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

При анализе данных, как правило, нет возможности рассмотреть всю интересующую нас совокупность объектов. Изучение очень больших объемов данных является дорогостоящим процессом, требующим больших временных затрат, к тому же неизбежно приводящим к ошибкам, связанным с человеческим фактором. Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть выборку, и получить интересующую нас информацию на ее основе. Однако размер выборки зависит от разнообразия объектов, представленных в генеральной совокупности. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности. 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.

Объекты

Многие инструменты DataMining при импорте данных из других источников предлагают выбрать тип шкалы для каждой переменной и/или выбрать тип данных для входных и выходных переменных (символьные и числовые, дискретные и непрерывные). Переменные могут являться числовыми данными либо символьными. Числовые данные, в свою очередь, могут быть дискретными и непрерывными. Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. Пример дискретных данных: продолжительность маршрута троллейбуса (количество вариантов продолжительности конечно): 10, 15, 25 минут. Непрерывные данные – данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность. Пример непрерывных данных: температура, высота, вес, длина и т.д. Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

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. Большинство инструментов DataMining позволяет импортировать данные из различных источников, а также экспортировать результирующие данные в различные форматы. Большинство BI-инструментов, представленных на рынке, использует слой метаданных или репозиторий. Метаданные (Metadata) – это данные о данных. В состав метаданных могут входить каталоги, справочники, реестры. Метаданные содержат сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Методы первичного анализа данных
Основы анализа данных

Описательная статистика, включающая технологии сбора и суммирования количественных данных, 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. 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 r, r □[0,1], 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:

Величина коэффициента корреляции, *r*	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-1
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными. Последовательность этапов регрессионного анализа:

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений;
2. Определение зависимых и независимых (объясняющих) переменных; 3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель;
3. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная);
4. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии);
5. Оценка точности регрессионного анализа;
6. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами.
7. Оценивается корректность и правдоподобие полученных результатов;
8. Предсказание неизвестных значений зависимой переменной. При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та часть, где оно меньше нуля, – к другому классу. Основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной.