

1 Введение

2 Data Mining как KDD

Источники [1] [2] [3]

1. Определение, данное на слайде часто цитируется во многих источниках. Оно было предложено в статье Fayyad et.al в 1996 году, но сам термин KDD появился на 7 лет раньше.
2. KDD - название, предложенное Григорием Пятецким-Шапиро для семинара, который он организовывал в рамках конференции IJCAI-89. Про воркшоп поговорим чуть позже в рамках исторической справки.
3. Хотя это определение часто цитируется в литературе, оно имеет ряд существенных неточностей.
4. Если интерпретировать valid, как “точный”, то зачастую возникает противоречие с остальными перечисленными качествами.
5. Рассмотрим, например, задачу анализа переписи населения. Одна из очень точных закономерностей, которую можно извлечь - женщины не служат в армии. Несмотря на высокую точность, у этой закономерности нет ни новизны, ни практической полезности.
6. Есть также и проблемы с интерпретируемостью закономерностей. В случае, когда решение принимается автоматической системой, она не нужна. А для человека интерпретируемость - слишком субъективное понятие.
7. Учитывая перечисленные недостатки данного определения, делаем вывод, что нам понадобится что-то лучше.

3 Data Mining как моделирование

Источники [4]

1. Более удобно рассматривать ДМ как процесс построения модели, хорошо описывающей данные. При этом можно формально определить модель и выбрать критерий, согласно которому можно утверждать, хорошая эта модель или плохая.
2. Выделяют следующие типы моделей
 - Статистические модели - те, в которых явно формулируется распределение вероятности, порождающее данные, возможно с набором неизвестных параметров. Алгоритм вычисления этих параметров называется алгоритмом обучения модели.
 - Подход, основанный на машинном обучении, характеризуется использованием одного из алгоритмов машинного обучения. При этом в основе алгоритма может лежать или не лежать статистическая модель. Наиболее удачно – когда мы не знаем о данных ничего.
 - Последнее время стал популярен вычислительный подход. Он подразумевает, что модель – это ответ на некоторый сложный запрос к данным. Такой подход позволяет делать наиболее слабые предположения о природе данных.

4 Пример 1. Красная икра на новогодний стол

Источники [5, ch. 2.3.4]

- Нужно купить красную икру, при этом важно не перепутать настоящую (дорогую) и искусственную (дешевую). Мы посмотрели в Интернете, сколько стоит икра в разных магазинах и идем выбирать в ближайший.

5 Некоторые важные события в истории Data Mining

Источники [1]

6 Data Mining – дисциплина тысячи имен

Источники [1]

1.

Список литературы

- [1] Journeys to Data Mining: The Journey of Knowledge Discovery
- [2] Knowledge Discovery in Databases: 10 years after
- [3] Are We Really Discovering “Interesting” Knowledge From Data?
- [4] Mining of Massive Datasets
- [5] Pattern Recognition and Machine Learning