



Лекция №14
ГЛУБИННЫЙ АНАЛИЗ ДАННЫХ

ТЕХНОЛОГИИ ХРАНИЛИЩ ДАННЫХ



Технология *Data Mining (DM)* предназначена для анализа структурированных данных с помощью математических моделей, основанных на статистических, вероятностных и оптимизационных методах, с целью выявления в них заранее неизвестных закономерностей, зависимостей и извлечения непредвиденной информации.

Основные задачи *DM*



- классификация;
- кластеризация;
- поиск ассоциаций и корреляций;
- выявление типовых образцов на заданном множестве;
- обнаружение объектов данных, не соответствующих установленным характеристикам и поведению;
- исследование тенденций во временных рядах и др.

ТЕХНОЛОГИИ ИАД

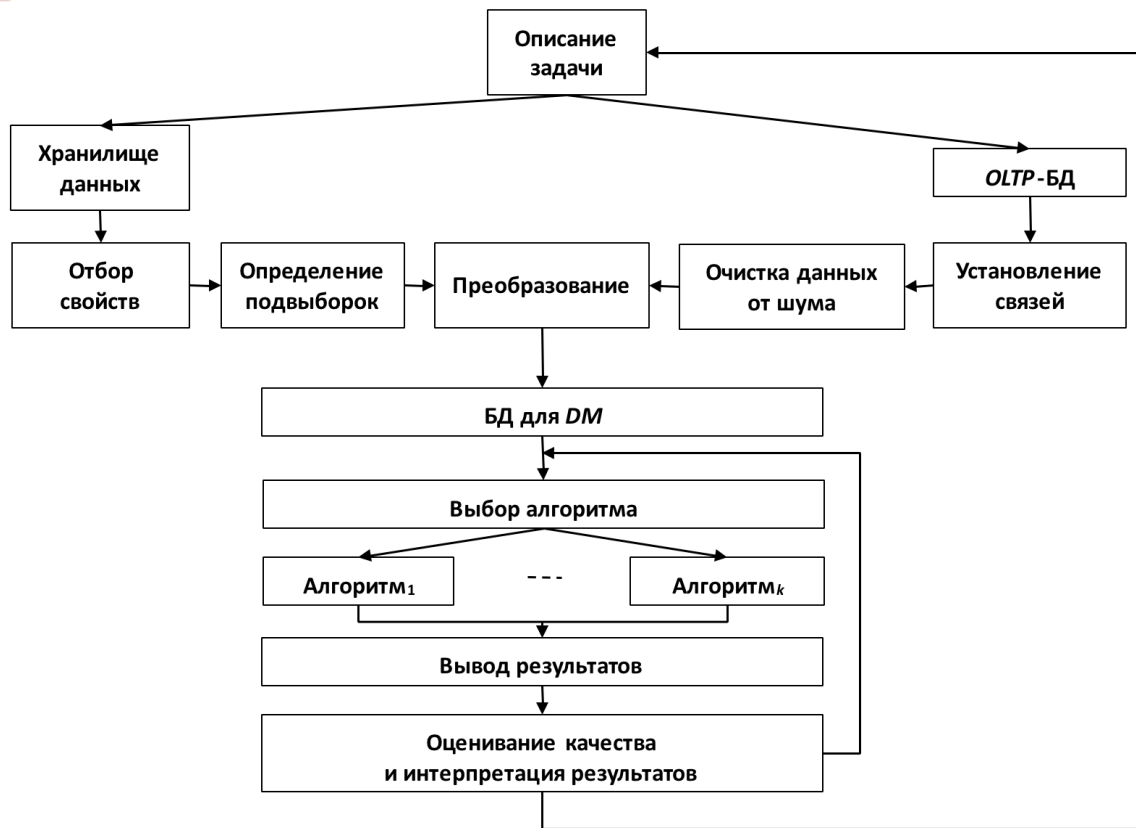


В рамках *DM* для сегментирования данных применяются **ИНС** и **методы кластерного анализа**.

Для индуктивного вывода — **деревья принятия решений**.

Для выявления в информационных массивах часто встречающихся пар объектов — **статистические и ассоциативные методы**.

Схема процесса ИАД на основе технологии *DM*



Процесс ИАД



Процесс ИАД включает четыре основных этапа:

1. На первом этапе аналитик формулирует постановку задачи в терминах целевых переменных;
2. На втором этапе осуществляется подготовка данных для анализа;
3. На третьем этапе проводится анализ данных с помощью методов *DM*;
4. На четвертом этапе осуществляется верификация и интерпретация полученных результатов (извлеченных знаний). При верификации применяется тестовый набор записей, выделенных из исходных данных и не подвергавшихся анализу.

Примеры продуктов *DM*



Примеры некоторых зарубежных продуктов *DM*:

1. *Intelligent Miner* (разработчик — фирма *IBM*). Используются ИНС, методы предсказывающего моделирования, обнаружения ассоциаций, сегментации БД и др.;
2. *Decision Series* (разработчик — *Neo Vista Software*). Используются ИНС, деревья и кластеры решений, ассоциативные правила;
3. *Darwin, Loyalty Stream* (разработчик — *Thinking Machines*). Используются ИНС и деревья решений.

В качестве примера российского продукта *DM* отметим систему *Poly-analyst* фирмы *Megaputer* (<http://www.megaputer.ru>).

Она позволяет выявлять многофакторные зависимости, которые представляются в виде функциональных выражений, а также формировать структурные и классификационные правила.

Примеры продуктов *DM*



В *Polyanalyst* используются:

- метод группировки и поиска ближайшего соседа;
- генетические алгоритмы;
- ИНС;
- статистические и ассоциативные методы;
- деревья решений;
- регрессионные модели;
- методы кластерного анализа;
- методы эволюционного программирования.

Унификация и стандартизация технологий *DM* являются целями проекта **CRISP-DM** — *Cross Industry Standard Process for Data Mining* (<http://www.crisp-dm.org>). Его результаты реализуются в рамках CASE-системы для разработки средств *DM*.