Вдовиченко І. Н., канд. техн. наук, доц.,
Криворізький національний університет

СТАТИСТИЧЕСКИЕ И АНАЛИТИЧЕСКИЕ СИСТЕМЫ, КАК КЛАССЫ ТЕХНОЛОГИИ DATA MINING

Технология Data Mining является новым витком в развитии средств и методов обработки данных. Data Mining переводится как «добыча» или «раскопка данных».

С развитием современных технологий, усовершенствованием аппаратного и программного обеспечения предприятий и организаций, на людей обрушились огромные потоки информационного «мусора».

Остро встал вопрос об эффективной переработке потоков сырых данных образующих информационную свалку. Как отмечает В.А. Дюк, специфика современных требований к такой переработке следующая:

- Данные имеют неограниченный объем.
- Данные являются разнородными (количественными, качественными, текстовыми).
- Результаты должны быть конкретны и понятны.
- Инструменты для обработки сырых данных должны быть прос-ты в использовании.

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных связей в данных. Эти шаблоны представляют собой закономерности, свойственные выборкам данных, которые могут быть выражены в специальной форме.

Построение и поиск шаблонов основаны на логике, выполняются методами, не зависящими от структуры и величины выборки значений анализируемых показателей. В них формируют не среднее значение, а общие закономерности. Для построения шаблонов могут использоваться выражения: встречаются ли, имеются ли, существуют ли. Ясно, что шаблоны должны быть не стандартны, особенны, и обеспечивать поиск таких данных и закономерностей, существование которых не предполагалось. Следовательно, технология Data Mining дает возможность получать новые знания из обычных информационных потоков, как сказано: «интеллектуальный анализ данных на основе открытий».

Data Mining является многоликой технологией, которая появилась и формируется на базе работ в областях прикладной статистики, теории баз данных и баз знаний, распознавания образов, аналитических методов и методов искусственного интеллекта, и др. Поэтому технология Data Mining в изобилии применяет комбинации методов и алгоритмов относящихся к различным научным и практическим сферам деятельнос-ти. Системы технологии интегрируют в себе одновременно несколько направлений.

Одним из классов систем, входящих в технологию Data Mining является класс статистических пакетов. Все современные статистические пакеты (наиболее мощные SAS, SPSS, STATGRAPICS, STATISTICA, STADIA и другие) наряду со статистическими методами имеют элементы Data Mining.

Математическая статистика долгие годы была основным инструментом анализа данных. Но при таких объемах информации (накопленная человечеством информация составляет 10^{24} байт =1 млн эксабайт = миллиард миллиардов байт), при традиционных методах обработки математическая статистика начала сдавать позиции. Основная причина — методики усреднения, в результате которых выполняются операции над фиктивными величинами. Методы математической статистики не заменимы для проверки гипотез и аналитической обработки. Но этого мало.

Поэтому, один из недостатков статистических пакетов, усложняющих их применение в Data Mining, является как раз усреднение характеристик выборки. Вследствие этого результаты корреляционного, регрессионного, факторного анализов и др. методов подвергаются оценке специальными шаблонами и наоборот. Диаграмма динамики представлена на рис.1.

Рис.1

Например, при выполнении научного исследования динамики гибели популяций животных по Украине использовали SPSS и соответст-вие шаблону «Существуют ли общие характеристики вымирающих видов и есть ли стереотип поведения у этих видов». В итоге получили следующие выводы. Прогнозируемый спад популяции: волков – 24 особи на 100 га за декаду, ежей ушастых – 2 особи на 100 га за декаду.