ВВЕДЕНИЕ

В современном развивающемся мире неуклонно растут объёмы накопленных данных. Одним из способов их переработки являются современные компьютерные технологии анализа данных, в том числе методы Data Mining в среде пакета STATISTICA [1].

Термин Data Mining можно перевести как «добыча данных» или «раскопка данных». Нередко под термином Data Mining подразумевают «обнаружение знаний в базах данных» (knowledge discovery in databases) и «интеллектуальный анализ данных» [2]. Вместе с определением добычи данных, также часто используется словосочетание «Data Warehousing» (хранилище данных). При этом понятие хранилища данных подразумевает способ хранения больших многомерных массивов данных, позволяющий легко извлекать информацию в процедурах анализа. Возникновение указанных терминов связано с дальнейшим развитием средств и методов обработки данных. В настоящее время термин Data mining — собирательное название, используемое для обозначения методов выявления в данных новых знаний, полезных и необходимых для принятия решений в различных сферах человеческой деятельности. Определение этому термину дал Григорий Пятецкий-Шапиро в 1996 году [3]: 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.

Совершенствование технологий сбора, хранения данных позволило накапливать огромные потоки информации в различных областях человеческой деятельности, поэтому применение статистических методов, основанных на парадигме среднего, стало неэффективным. Одна из причин та, что в основе прикладной статистики лежат операции над виртуальными величинами (средняя температура больных в больнице, средняя зарплата граждан по стране и т. д.). Другая причина в структуре современных данных, которые могут быть разнородными (количе-

ственными, качественными, текстовыми) и неограниченного объема.

С помощью Data Mining можно обработать исходные данные путем запуска автоматизированного поиска закономерностей (паттернов). Данные шаблоны чаще всего ищутся среди фрагментов неоднородных многомерных данных. Процесс добычи данных подразумевает под собой три последовательных этапа выполнения: исследование данных, построение модели и ее проверку. В Data Mining гипотезы формулируются без участия человека, равно как и ищутся необычные (unexpected) шаблоны, чего не скажешь о традиционно используемой оперативной аналитической обработке данных OLAP (online analytical processing). OLAP — технология оперативной аналитической обработки данных, использующая методы и средства для сбора, хранения и анализа многомерных данных в целях поддержки процессов принятия решений [3].

Различают пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining: ассоциация, последовательность, классификация, кластеризация и прогнозирование [2]. Ассоциация имеет место в том случае, если несколько событий связаны друг с другом. Если существует цепочка связанных во времени событий, то говорят о последовательности. С помощью классификации выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил, позволяющих предсказать принадлежность объектов к той или иной группе. Кластеризация отличается от классификации тем, что сами однородные группы (кластеры) сходных (похожих) между собою объектов заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные кластеры заданной совокупности данных. Основой для всевозможных систем прогнозирования служит историческая информация, представленная в виде временных рядов. Если удается найти модели, адекватно отражающие динамику поведения целевых показателей, то с их помощью можно предсказать и поведение системы в будущем.

В настоящем пособии рассмотрены методы машинного обучения Data Mining применительно к решению задач классификации, кластеризации, прогнозирования.