XXIX Неделя науки СПбГТУ. Материалы межвузовской научной конференции. Ч.V: С.71-72, 2001.

УДК 007.52

А.А. Аккуратов (6 курс, каф. САиУ), Л.А. Станкевич, к.т.н., доц.

ВЫЧИСЛЕНИЕ ФУНКЦИЙ РАСКОПКИ ДАННЫХ (DATA MINING) С ПРИМЕНЕНИЕМ КОГНИТИВНЫХ МОДУЛЕЙ

Раскопка данных (DM - Data Mining) является процессом обработки потоков данных с целью эффективного извлечения информации, полезной для принятия решений в определенной проблемной области. ДМ относится к области раскрытия знаний из баз данных (Knowledge Discovery).

Идея ДМ, как экстракции ценной информации из данных, не нова. Новым является использование специальных компьютерных технологий для реализации ДМ в различных проблемных областях, позволяющие Гига - и Терра-байтные данные обрабатывать в on-line режиме (в реальном времени) и использовать их при принятии решений. Только сейчас специалисты осознали, что методами ДМ можно переработать огромные потоки "информационной руды", которые иначе могут просто превратиться в "свалку". Доступность и удобство работы с современными программными инструментами ДМ также придали популярность этому виду обработки данных. ДМ можно использовать везде, где есть данные. Отдача от ДМ может достигать 1000%. Наибольшее применение ДМ находит в коммерческой сфере.

Цель ДМ - выявление скрытых правил и закономерностей в больших наборах данных. Человек плохо воспринимает большие массивы данных и не может уловить более трех взаимодействий в данных. Традиционная математическая статистика также пасует в решении реальных задач обработки данных. Она оперирует средними характеристиками выборки часто являющимися фиктивными величинами (например, среднее число пассажиров на станциях). Поэтому такие методы полезны в основном для проверки заранее сформулированных гипотез (verification-driven DM). Современная техника ДМ (discovery-driven DM) обрабатывает данные с целью автоматического поиска патернов, характерных для фрагментов неоднородных многомерных данных. В отличие от OLAP (OnLine Analytical Processing) в ДМ формулировка гипотез и выявление необычных образов переложено с человека на ЭВМ.

Существует много различных функций ДМ среди которых наиболее часто используются функции такие, как: фильтрация, компрессия, анализ главных компонент, кластеризация, ассоциация, классификация, моделирование, последовательное разделение образов, предсказание временных серий, раскопка закономерностей. Диапазон техник для реализации этих функций включает: статистики, грубые множества, нейронные сети и пр.

В данной работе предлагается вычислять перечисленные функции ДМ с использованием специально разработанных когнитивных нейрологических модулей (обучаемых и самообучаемых).

Базовый нейрологический модуль (NM) достаточно легко реализуется в программных вариантах. Одним из вариантов для вычисления функций ДМ может быть предлагаемый нечетко-логический вариант NM.

Достоинствами NM являются: быстрое обучение, простота реализации и хорошие аппроксимирующие свойства. Это позволяет более эффективно, чем нейронные сети, использовать его для реализации многих функций ДМ. Наиболее успешно реализуются функции: ассоциации, классификации, моделирования и раскопки знаний. Недостатком NM является значительный расход памяти на хранение обработанной информации. Поэтому существуют некоторые проблемы при реализации функций: фильтрации, кластеризации, последовательных образов и прогнозирования временных серий, которые требуют обработки большого количества данных. В этом случае предполагается ввести порционную обработку данных и после обработки каждой порции производить сжатие запомненной информации с помощью хеширования.

При решении задач прогнозирования обычно требуется реализовать несколько функций ДМ. Для этого можно построить систему из нескольких NM, каждый из которых исполняет отдельную функцию ДМ, а вместе они выполняют процесс, в рамках которого решается задача в целом.

Процесс решения задачи прогнозирования, основанный на применении функций ДМ, включает четыре этапа: подготовка данных, предварительная обработка данных, обработка функциями ДМ, анализ результатов. Система для решения задачи состоит из нескольких уровней: нижнего - процедур подготовки данных (контроль, нормализация, шкалирование), реализуемых традиционными методами; среднего 1 - функций предварительного процессирования данных (очистки, сегментации, препроцессирования), реализуемых нейросетевыми модулями; среднего 2 - функций основной обработки данных (кластеризации, последовательных образов, прогнозирования временных серий, ассоциации), реализуемых NM автоматически; верхнего - функций анализа результатов (анализ на правилах, визуальный анализ), реализуемыми NM под контролем оператора.

Предполагается, что перечисленные функции, начиная со средних уровней, формируются в NM путем обучения по примерам или самообучения. Обученная система позволяет решать задачи прогнозирования в реальном времени даже при больших потоках сырых входных данных.

Рассмотренный вариант системы с NM, настроенными путем обучения на вычисление функций ДМ, опробован на ряде задач прогнозирования из области финансового менеджмента, маркетинга, транспортных систем и пр. Предполагается развить этот подход в рамках многоагентных нейрологических систем, позволяющих решать сложные задачи планирования.