Известия Гомельского государственного университета
имени Ф. Скорины, №5(50), Ч1, 2008

УДК 004.02 : 004.9

О структуре программно-технического комплекса анализа многомерных объектов

Ю. В. Гуща, А. И. Жукевич, В. Г. Родченко

Введение. Анализ данных является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, теории распознавания образов, методов искусственного интеллекта, теории баз данных. Системы анализа данных применяются как для решения разнообразных прикладных задач, так и в качестве инструментария для проведения уникальных исследований (генетика, химия, медицина и др.). В настоящее время количество инсталляций подобных систем в виде соответствующего программного обеспечения достигает многих десятков тысяч.

В начале 90-х годов прошлого столетия рынок систем анализа данных насчитывал порядка десяти основных поставщиков. В основном речь шла об универсальных статистических пакетах, основанных на методах прикладной статистики, например: STADIA, SYSTAT, STATGRAPHICS, SPSS и специализированные пакеты: Эвриста, Мезозавр, КЛАСС-МАСТЕР, САНИ [1, 2]. К стредине 90-х число поставщиков, представленных компаниями малого, среднего и большого размера, насчитывало более пятидесяти фирм. По данным результата опроса “Инструменты Data Mining, которые Вы регулярно используете”, проведенного в мае 2007 года на Kdnuggets (http://www.kdnuggets.com) наибольшее количество голосов получили среди коммерческих продуктов: SPSS Clementine, Salford CART/MARS/TreeNet/RF, Excel, SPSS, SAS, Angoss, KXEN. Среди свободно распространяемых продуктов лидируют Yale, Weka и R, и около 11% опрошенных сообщили, что используют продукты собственной разработки.

Имеется большое количество различных систем с развитым набором методов и алгоритмов анализа данных, причем многие из таких систем интегрируют в себе сразу несколько подходов. В то же время, как правило, в каждой системе присутствует какая-то ключевая методика исследования, которая является доминирующей. Учитывая соответствующие методы исследования данных, выделяют следующие виды систем: статистические пакеты, нейронные сети, предметно-ориентированные аналитические системы, системы рассуждений на основе аналогичных случаев, на использовании алгоритмов ограниченного перебора, на основе генетических алгоритмов, деревьев решений, эволюционного программирования [3].

В состав практически всех наиболее распространенных статистических пакетов (SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA) включаются как традиционные статистические методы, так и элементы анализа данных. Основное внимание в них в первую очередь уделяется корреляционному, регрессионному, факторному анализу. К основным недостаткам этих пакетов относятся достаточно высокий уровень требований к специальной подготовке пользователя и “тяжеловесность” для массового применения при решении конкретных прикладных и исследовательских задач.

Нейронные сети представляют собой системы, архитектура которых устроена по принципу нервной ткани из нейронов. Перед непосредственным применением сеть должна быть натренирована (обучена) с использованием предварительно полученных данных, для которых известны как значения входных параметров, так и правильные ответы на них. Обучение в данном случае фактически предполагает подбор весов межнейронных связей, которые обеспечивают наибольшую близость ответов сети к априори известным правильным ответам. Одним из существенных недостатков технологии нейронных сетей является необходимость в обеспечении большого объема обучающей выборки. Еще один недостаток заклю-

чается в том, что натренированная нейронная сеть по своей сути представляет собой черный ящик. Знания, формализованные через фиксированные веса огромного числа межнейронных связей, практически не поддаются анализу и интерпретации.

В области исследования финансовых рынков широкое распространение получили предметно-ориентированные аналитические системы, которые еще называют системами “технического анализа”. Основанные на различных эмпирических моделях динамики рынка, такие системы интегрируют в себе совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля. Применяемые методы ориентированы на использование несложного статистического аппарата, но, как правило, максимально учитывают сложившуюся специфику в соответствующей предметной области (профессиональный язык, системы различных индексов и т.п.).

При решении разнообразных прикладных задач эффективным оказывается использование систем рассуждений на основе аналогичных случаев (CBR - case based reasoning), когда реализуется выбор близкого аналога для исходных данных из уже имеющихся исторических данных. Метод, используемый в данных системах, также называют методом «ближайшего соседа». Существенным недостатком подобных CBR-систем является то, что они при поиске решения не ориентированы на построение какого-либо набора решающих правил, обобщающих предыдущий опыт, а в выборе решения основываются на всем массиве доступных исторических данных. В данном случае невозможно конкретизировать, на основе чего система строит свои ответы. Еще один недостаток проявляется в произвольном выборе в CBR-системе соответствующей “меры близости”, хотя от этого выбора существенно зависит объем множества прецедентов, которые необходимо напрямую учитывать при проведении удовлетворительной классификации или при формировании прогноза [4].

В середине 60-х годов М.М. Бонгардом были разработаны алгоритмы ограниченного перебора, которые ориентированы на выявление логических закономерностей в данных. Этими алгоритмами предусматривается вычисление частот комбинаций простых логических событий в подгруппах данных. По результатам анализа вычисленных частот принимается заключение о полезности той или иной комбинации для проведения классификации. Указанный подход реализован в системе WizWhy (фирма WizSoft), которая по некоторым оценкам является на сегодняшний день одним из лидеров на рынке продуктов анализа данных и которая постоянно демонстрирует более высокие показатели при решении практических задач в сравнении с другими алгоритмами.

Генетические алгоритмы в первую очередь направлены на решение разнообразных комбинаторных задач и задач оптимизации, однако в последнее время они нашли свое применение и в качестве инструментария в задачах анализа данных. Удобство генетических алгоритмов проявляется в том, что они позволяют легко проводить процедуру распараллеливания, например, можно разбить поколение на группы и работать с каждой из них независимо, обмениваясь, время от времени? несколькими хромосомами. Недостатком генетических алгоритмов является то, что критерий отбора хромосом и используемые процедуры являются эвристическими и не гарантируют нахождения “лучшего решения”. Как и в реальной жизни, эволюцию может “заклинить” на какой-либо непродуктивной ветви. Кроме того, можно привести пример, когда два неперспективных родителя, которые исключаются из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Указанные недостатки особенно начинают проявляться при решении высокоразмерных задач со сложными внутренними связями.

Используется подход с применением так называемых деревьев решений (decision trees), представляющих собой иерархическую структуру, базирующуюся на наборе вопросов, подразумевающих ответ “да” или нет”. Хотя такой способ обработки данных далеко не идеально находит существующие закономерности, он довольно часто используется в системах прогнозирования в силу наглядности получаемого ответа. Соответствующим алгоритмом реализуются простейший способ последовательного просмотра признаков с целью построения логического вывода. К наиболее известным системам, в которых присутствует реализация этого метода, относятся See5/C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Be-

ликобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада).

Эволюционное программирование ориентировано на поиск и генерацию алгоритма, выражающего взаимозависимость данных. Этот процесс может осуществляться либо на основе изначально заданного алгоритма, модифицируемого в процессе поиска, либо иногда поиск взаимозависимостей производится среди каких-либо предварительно заданных видов функций (например, полиномов). Данный подход используется в системах PolyAnalyst, NeuroShell (Ward Systems Group).

Несмотря на обилие методов анализа данных, приоритет постепенно все более смещается в сторону логических алгоритмов поиска использующих if-then правила. С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации базы данных, извлечения из данных “скрытых” знаний, интерпретации данных, установления ассоциаций в базе данных и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

Вместе с тем, главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор (алгоритмы KOPA, WizWhy) [5], либо строят деревья решений (алгоритмы CART, CHAID, ID3, See5, Sipina и др.), имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил.

При проведении научных исследований и решений целого ряда задач, связанных с анализом многомерных объектов сложной природы, высокую эффективность продемонстрировали подходы, базирующие на использовании методов математической теории распознавания образов [6]. Процесс распознавания реализуется через выполнение двух основных процедур, первая из которых ориентирована на *обучение*, а вторая – непосредственно на *распознавание* [7]. Обучение осуществляется на основе анализа данных классифицированной обучающей выборки и в результате формируются эталоны классов. Эти эталоны в дальнейшем используются при выполнении непосредственно процедуры распознавания исследуемых объектов сложной природы.

Для реализации процедуры распознавания образов многомерных объектов с помощью указанного метода автоматизации анализа данных предлагается разработать специальный программно-технический комплекс анализа многомерных объектов ПТКАМО.

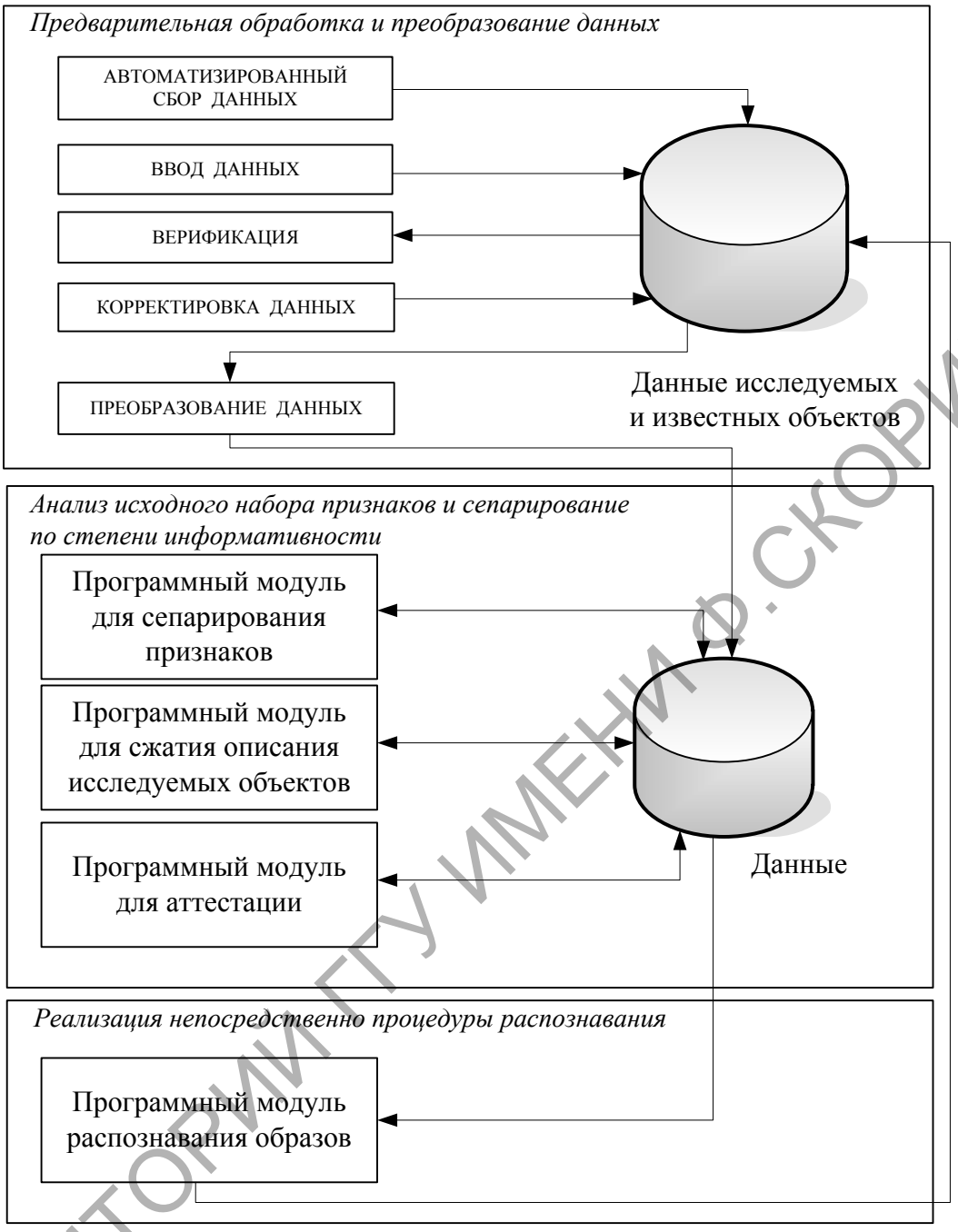
Структура программно-технологического комплекса. Структурно пакет ПТКАМО включает в себя три функциональные части: 1) предварительной обработки и преобразования данных; 2) анализа исходного набора признаков и сепарирования их по степени информативности; 3) реализации непосредственно процедуры распознавания. Структурная схема связей модулей ПТКАМО в виде группы технологических переходов обработки данных представлена на рисунке.

Первая часть программного комплекса ПТКАМО является проблемно-ориентироВанной и состоит из программ, которые, во-первых, позволяют осуществить автоматизированный сбор данных (при наличии возможностей организации автоматизированного сбора) и ввод данных для сохранения их в базе данных или в виде текстовых файлов, произвести верификацию введенной информации и провести при необходимости корректировку, во-вторых, реализовать необходимые операции по преобразованию исходных данных к матричному виду типа “объект-свойство” в универсальный формат представления. Процесс автоматизированного сбора информации предусматривает применение как программных, так и аппаратных средств.

Вторая часть комплекса предназначена для анализа исходного набора признаков и сепарирования их по степени информативности.

В состав второй части входят:

- программный модуль для сепарирования признаков из исходного априорного словаря признаков на три вида по степени их информативности;



Структурная схема связей модулей ПТКАМО.

- программный модуль для сжатия описания исследуемых объектов в уточненном признаковом пространстве и проведения нормировки значений признаков;
- программный модуль для аттестации возможности использования построенного уточненного словаря признаков для анализа данных.

Третья часть программно-технического комплекса состоит из программного модуля, в котором реализован алгоритм метода автоматизации анализа данных, предусматривающий выполнение непосредственно процедуры распознавания исследуемого объекта.

Заключение. Разработана структурная схема программно-технического комплекса анализа многомерных объектов, который базируется на использовании методов теории распознавания образов и кластерного анализа. Для качественной реализации процедуры распознавания предусматривается обязательное выполнение процедуры обучения, которая осуществляется на основе анализа данных классифицированной обучающей выборки. Признаки из исходного априорного словаря сепарируются по степени информативности с точки зрения

разделения эталонов исследуемых классов в соответствующем многомерном признаковом пространстве решений.

Метод построения комплекса предусматривает автоматизацию процедур обучения и распознавания исследуемых объектов. Он характеризуется универсальностью и позволяет исследовать объекты на основе анализа различных по своей природе исходных признаков.

Abstract. Processing big volumes of the data is impossible without the application of modern computing technologies. There are various systems based on different methods and algorithms of data analysis. Each system is classified according to the presence of the certain methods of data analysis. For the realization of analysis algorithm of the given multivariate objects method the related approach is offered in the paper. The algorithm is based on the methods of the mathematical pattern recognition theory.

Литература

1. С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин, Прикладная статистика: Классификация и снижение размерности, Москва, Финансы и статистика, 1989.

2. Ю.Н. Тюрин, Анализ данных на компьютере, Москва, Финансы и статистика, 1995.

3. М. Киселев, Е.Соломатин, Средства добычи знаний в бизнесе и финансах, Открытые системы, № 4 (1997), 41-44.

4. В.А. Дюк, Обработка данных на ПК в примерах, Санкт-Петербург, «Питер», 1997.

5. Ю.И. Журавлев, Распознавание, классификация, прогноз. Математические методы и их применение, Москва, Наука, Выпуск.2, 1989.

6. Н.Г. Загоруйко, Прикладные методы анализа данных и знаний, Новосибирск, Издательство Института математики СО РАН, 1999.

7. В.Г. Родченко, Об одном методе построения системы распознавания образов, Известия Гомельского государственного университета имени Ф.Скорины, №6 (2002), 93-96.

Гродненский государственный
университет имени Янки Купалы

Поступило 13.05.08