УДК 62-50

ПРОГРАММНО-АЛГОРИТМИЧЕСКИЙ КОМПЛЕКС СТРУКТУРНО-КЛАССИФИКАЦИОННОГО АНАЛИЗА СЛОЖНООРГАНИЗОВАННЫХ ДАННЫХ [1]

© Бауман Е.В., Дорофеюк А.А., Дорофеюк Ю.А., Киселёва Н.Е.

Институт проблем управления РАН

Abstract. In this paper the conception of the structure-ranging analysis methods application is discussed, among them were discussed such methods as: parameter structuring; objects classification; dynamic ranging analysis; piecewise approximation of complex dependences. This conception is realized in the man-machine system with intellectual interface for user. During the complex development a special attention was given to problems, in which researched objects have clear-cut areal structure.

Введение

В работах [1-4] предложены принципы построения и использования алгоритмов структурного анализа сложноорганизованных данных. В работе предлагается концепция применения разнообразных методов структурно-классификационного анализа, включающая методы: структуризации набора исходных параметров; структуризации множества объектов с помощью алгоритмов автоматической классификации; динамического классификационного анализа, позволяющие анализировать поведение объектов в многомерном пространстве траекторий; анализа сложных, нелинейных зависимостей с помощью алгоритмов кусочной аппроксимации. Реализация этой концепции подразумевает создание человеко-машинной системы с интеллектуальным интерфейсом для пользователя – предметника, позволяющая формировать пользователем схемы использования алгоритмов, текущего анализа промежуточных результатов, наглядного их отображения.

1. Постановка задачи классификационного анализа данных

Структура исходных данных в задачах классификационного анализа. Функционирование любой сложной системы описывается состоянием составляющих её элементов (объектов) и их взаимодействием. Соответственно, данные о системе представляют собой либо таблицу значений некоторых параметров, характеризующих состояние объектов, либо таблицу, отражающую взаимодействие между объектами, либо, наконец, таблицу связи между параметрами. При описании концепции системы, опирающейся на методологию классификационного анализа, ограничимся таблицами первого типа: «объекты-параметры». Заметим, что данные о системе обычно фиксируются ни в один момент времени, а многократно в течение некоторого периода (например, ежемесячно, ежеквартально или ежегодно в течение несколько лет работы системы). Характерной особенностью исследования реальных систем является невозможность получения значений всех параметров по всем объектам во все моменты времени, что приводит к пропускам в данных.

[1]Работа выполнена при частичной финансовой поддержке РФФИ, проекты 08-07-00349-а, 08-07-00347-а.

Итак, исходный материал о функционировании исследуемой системы представляет собой куб данных «объекты-параметры-время» $\left\| x_i^{(j)(t)};\ i=\overline{1,n};\ j=\overline{1,k};\ t=\overline{1,T} \right\|$, причем в данных возможны пропуски (здесь $x_i^{(j)(t)}$- значение j-го параметра на i-м объекте в момент времени t) .

Основная цель классификационного анализа заключается в выявлении наиболее общих закономерностей функционирования анализируемой системы: структуризация набора параметров для выявления групп тесно связанных параметров и построения небольшого числа интегральных показателей; структуризация множества объектов, для чего необходимо выделить в пространстве выбранных параметров области, отражающие типовые режимы деятельности отдельных объектов системы; построение кусочных моделей зависимости некоторых выходных параметров от входных; выявление динамических свойств системы – выделение характерных траекторий изменения параметров во времени, выявление зависимостей между параметрами с учетом временного сдвига и т.д.

2. Состав программно-алгоритмического комплекса (ПАК)

Для реализации вышеперечисленных целей был разработан человеко-машинный программно-алгоритмический комплекс (ПАК) с интеллектуальным интерфейсом для пользователя-предметника. В него входит база данных, в которой хранится исходный куб данных, наименования объектов, названия параметров, результаты всех этапов обработки. Для реализации алгоритмов структурно-классификационного анализа данных ПАК содержит 5 основных обработочных модулей**:** предварительной обработки и фильтрации исходных данных, классификационного анализа параметров, классификации объектов, анализа множества полученных классификаций, кусочной аппроксимации. ПАК позволяет постоянно обновлять куб данных, при этом все полученные ранее результаты классификационного анализа распространяются на новые данные. ПАК оснащён дружественным интерфейсом, который включает структуру меню для выбора режимов обработки, вводные формы для определения свободных параметров и т.д. ПАК позволяет отображать исходные данные и результаты анализа в наглядной форме, в том числе в виде географической карты, гистограмм, графиков и т.д.

Работа с комплексом организована в виде диалога. На каждом этапе пользователю предоставляется возможность выбрать один из основных модулей обработки. В то же время ему даётся рекомендация, - какую обработку целесообразно проводить на данном этапе. Результаты применения программ каждого модуля заносятся в базу данных и являются исходными данными для работы других модулей.

На рис. 1. представлена общая блок-схема ПАК, отражающая рекомендуемую последовательность применения основных модулей (блоков). Ниже описывается каждый из основных блоков отдельно.

2.1. **Предобработка.** До проведения структурно-классификационного анализа необходима предварительная обработка: статистический анализ, выявление грубых ошибок в данных, заполнение пропусков в данных и т.п.

Рис. 1. Блок-схема ПРОГРАММНО-АЛГОРИТМИЧЕСКОГО КОМПЛЕКСА

Как для статистического, так и для структурно-классификационного анализа стандартным видом данных являются таблицы типа «объекты-параметры».

Существует два способа развертки исходного куба данных в таблицу такого рода. Первый способ заключается в том, что в качестве объектов таблицы рассматриваются пары «объект - момент времени» исходного куба данных (например, данные об одном и том же объекте за t лет рассматриваются как данные о t разных объектах). Набор параметров при этом остается без изменений. Такую развертку куба будем обозначать через $\mathrm{T}^{\text{пар}}_{\text{об-вр}}$. При втором способе множество объектов таблицы совпадает с множеством объектов исходного куба данных, а в качестве параметров

рассматриваются пары «параметр - момент времени». Соответствующая развертка обозначается как $T_{об}^{пар\text{-}вр}$.(см. рис.2).

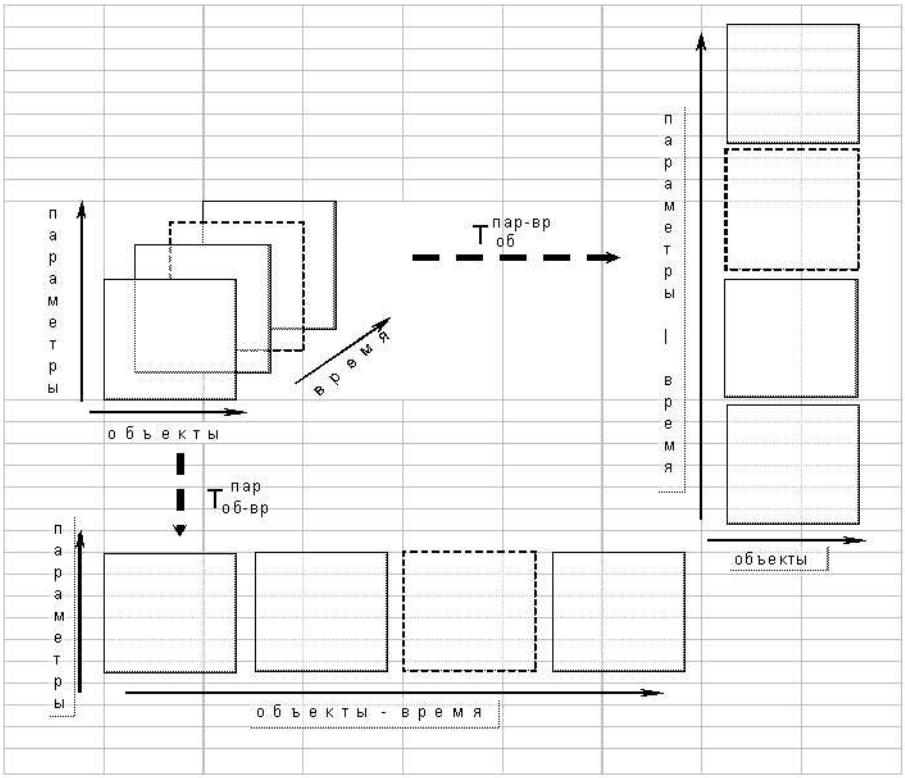

Рис. 2. Варианты развёртки куба данных

С помощью статистического анализа определяются средние значения параметров, их дисперсии и корреляции между ними. Такой анализ позволяет оценивать как независящие от времени статистические характеристики параметров, так и связи между параметрами, учитывая сдвиги во времени. Далее выявляются грубые ошибки в данных, т.е. значения параметров, сильно отличающиеся от средних значений.

Алгоритмы классификационного анализа модифицированы так, чтобы они работали и на данных с пропусками. С другой стороны, часто удобно заполнить пропуски до начала основного классификационного анализа.

Для заполнения пропусков предлагается следующая процедура:

Сначала производится группировка параметров, а затем по каждой группе параметров делается классификация объектов. Пропущенное значение $x_i^{(j)}$ таблицы заполняется следующим образом: находится группа, которой принадлежит j-й параметр; в соответствующей классификации находится класс, в котором лежит i-й объект, и в качестве $x_i^{(j)}$ берется среднее значение j-ого параметра по этому классу.

2.2. **Структуризация исходных параметров.** Опыт использования алгоритмов структурно-классификационного анализа показывает, что классификация по всем исходным параметрам не всегда приводит к желаемым результатам. Действительно,

при сравнительно небольших выборках экспериментальных наблюдений и наличии помех (ошибки в определении значений параметров, сознательное искажение информации и т.д.) использование для классификации большого числа входных параметров приводит к сильному «перемешиванию» классов, а сами классы при этом плохо поддаются интерпретации. По этой причине классификацию целесообразно проводить не в исходном пространстве, а в пространстве наиболее существенных (информативных) параметров, имеющем значительно меньшую размерность. Для выявления структуры исходного набора параметров вне зависимости от времени используется развертка исходного куба данных в таблицу типа $\mathrm{T}^{\text{пар}}_{\text{об-вр}}$. Структуризация проводится с помощью алгоритмов экстремальной группировки параметров [1],[5]. Определяется, нужна ли группировка с фоновой группой или без неё (т.е. отсекать или нет сильно шумящие параметры). Результатом группировки будут группы параметров и факторы – обобщённые характеристики групп. На основе результатов группировки строятся интегральные показатели исследуемой системы. В качестве таковых выбираются либо сами факторы, либо параметры в определённом смысле ближайшие к факторам. Основное условие – они должны быть легко интерпретируемы. Для удобства использования интегральных показателей по каждому из них делается одномерная классификация объектов. Благодаря этому интегральный показатель преобразуется в качественный, так как его значения можно качественно характеризовать в терминах типа «низкие», «средние», «высокие».

Другое применение метода экстремальной группировки – выбор информативных параметров для других методов анализа. Выбирается либо набор факторов, либо набор, в который входят по одному или небольшое число параметров из каждой группы. Обычно окончательное решение о выборе информативных параметров производится экспертом-пользователем.

Для определения зависимостей между параметрами с учётом времени, используется развертка куба данных, в которой за параметры принимаются пары «параметр-момент времени», т.е. рассматривается таблица вида $\mathrm{T}^{\text{пар-вр}}_{\text{об}}$.

2.3. **Классификация объектов.** Ключевым этапом структурного анализа исследуемой системы является структуризация множества элементов системы (объектов). В ПАК для решения этой задачи предназначены алгоритмы автоматической классификации. При этом используется целый ряд проблемно ориентированных алгоритмов – в детерминированной и размытой постановке, с фоновым классом, итерационных (на каждом шаге обрабатывается только один объект выборки), параллельных (когда все объекты выборки обрабатываются одновременно) и т.д. [1].

Вначале производится выбор пространства показателей, которое является исходным для классификации объектов.

Для нахождения независимых от времени режимов функционирования объектов исследуется таблица, у которой набор параметров совпадает с исходным, а роль объекта играет пара «объект-момент времени» (таблица $\mathrm{T}^{\text{пар}}_{\text{об-вр}}$). Классификация объектов из таблицы $\mathrm{T}^{\text{пар-вр}}_{\text{об}}$, у которой множество объектов совпадает с исходным,

а в качестве параметров рассматриваются пары «параметр-момент времени», приводит к нахождению типичных траекторий изменения параметров во времени. Для структуризации объектов изменяющихся во времени используются методы динамического классификационного анализа, позволяющие анализировать поведение объектов в многомерном пространстве траекторий. Здесь также представлен достаточно широкий спектр проблемно ориентированных алгоритмов – по типу обрабатываемых траекторий, по используемым мерам связи (близости) траекторий и т. д. [2].

До применения выбранного алгоритма классификации необходимо определить: вид функционала; строится ли классификация с фоновым классом или без (т. е. отбрасываются ли далекие объекты); тип размытости**:** четкая, размытая, с размытыми границами, четкая с размытым фоном, размытая с четким фоном и т. д. [1].

Результатом классификации являются функции принадлежности объектов к классам и описание самих классов.

2.4. **Структуризация результатов классификации.** Практически все алгоритмы структурно-классификационного анализа содержат свободные параметры, значения которых трудно выбрать заранее из теоретических соображений. Кроме того, эти алгоритмы находят лишь локальный экстремум соответствующего критерия, поэтому результаты их работы зависят от начальных условий (начального разбиения объектов на классы и параметров на группы). В связи с этим, при решении практических задач свободные параметры алгоритмов, начальные условия, а часто и состав переменных, образующих исходное пространство, варьируются в широких пределах. Это приводит к тому, что образуется целое множество различных вариантов классификации. Число классификаций часто оказывается столь большим, что для их анализа приходится применять машинные методы, вводя меру близости между классификациями и разбивая их на группы «похожих» классификаций. Одна из таких мер предложена в [6]. Применение компьютерных методов обработки результатов классификации существенно облегчает их дальнейший неформальный анализ.

2.5. **Кусочная аппроксимация.** После структуризации исходных данных как по параметрам, так и по объектам можно, в случае необходимости, приступить к построению моделей зависимости выходных показателей от входных. Важной частью ПАК является модуль анализа сложных, нелинейных зависимостей и формирования математических моделей различных процессов.

Для этой цели используются алгоритмы кусочной аппроксимации [1],[4]. Предусмотрен специальный режим одномерной кусочной аппроксимации, когда соответствующий алгоритм обеспечивает получение глобально-оптимального решения задачи [4]. Классы объектов, полученные автоматической классификацией, используются либо непосредственно как области действия отдельных локальных моделей (при двухступенчатом методе аппроксимации), либо как начальные условия для алгоритма нахождения таких областей (при одноступенчатом алгоритме аппроксимации). Обычно для кусочной аппроксимации используется первый вариант развертки исходного куба данных, т.е. в качестве объектов рассматриваются пары «объект-момент времени». Иногда бывает удобно для таких пар учитывать время в качестве входного

параметра при построении кусочной модели. Для построения кусочной аппроксимации необходимо определить: выходной параметр; пространство входных параметров; число классов; нужен ли фоновый класс; тип размытости классификации [4].

Заключение

Разработанный программно-алгоритмический комплекс предназначен для решения задач анализа сложноорганизованных многомерных данных, а также для поддержки принятия решений при анализе и реформировании крупномасштабных систем управления. Предусмотрена возможность эксплуатация системы пользователями двух уровней. На первом уровне пользователь-аналитик формирует модели исследуемой системы, в том числе: набор интегральных показателей, пространство, в котором проводится классификация, результирующая классификация, результирующие кусочные прогностические модели. На втором уровне пользователь-предметник использует полученные на первом уровне модели для решения задач оперативного управления.

Разработанный ПАК использовался для решения многих прикладных задач, в том числе в задачах управления региональным здравоохранением [7], региональными пассажирскими автоперевозками [8], а также при обработке сложноорганизованных данных (например, при обработке пульсовых сигналов лучевой артерии [9]).

Список литературы

1. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных // Избранные труды Международной конференции по проблемам управления. Том 1. / – М. : СИНТЕГ. 1999. – С. 62-67.
2. *Чернявский А.Л., Бауман Е.В., Дорофеюк А.А.* Методы динамического классификационного анализа данных // Искусственный интеллект, № 2. 2002. – С. 290-298.
3. *Бауман Е.В., Дорофеюк А.А., Чернявский А.Л.* Методы структурной обработки эмпирических данных // Измерения, контроль, автоматизация. 1985. № 3. - С. 64-69.
4. *Бауман Е.В., Дорофеюк А.А., Корнилов Г.В.* Алгоритмы оптимальной кусочно-линейной аппроксимации сложных зависимостей // Автоматика и телемеханика. 2004. № 10. – С. 163-171.
5. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных // – М.: Наука. 1983. – 464 с.
6. *Бауман Е.В.* Структуризация номинальных признаков в задачах экспертизы. – В кн.: Экспертные оценки в задачах управления. Сборник трудов. // – М.: Институт проблем управления. 1982. – С. 22-26.
7. *Бауман Е.В., Дорофеюк А.А., Чернявский А.Л., Медик В.А.* Классификационные методы в аналитических задачах регионального управления // Труды Института проблем управления РАН. Том X. // – М.: ИПУ РАН. 2000. – С. 38-40.
8. *Блудян Н.О., Чернявский А.Л.* Структурные методы совершенствования управления региональным пассажирским автотранспортом // М.: «Альфа-Мир». Серия Транспорт. 2002. – 127 с.
9. *А.А.Десова, А.А.Дорофеюк, В.В.Гучук, Ю.А.Дорофеюк, И.В.Покровская* Процедуры классификационного анализа в задаче формирования информативных признаков при исследовании ритмической структуры биосигнала. // Автоматика и телемеханика. 2008. № 6. – с. 143-152.

Статья поступила в редакцию 28.04.2008