ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)
1984 год Выпуск 102

УДК 62-5:621.391

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ РАСПОЗНАВАНИЯ И КЛАССИФИКАЦИИ КОД-2

К.Е.Волковицкий, Е.М.Крылов, И.Б. Сиродка,
В.А.Дискант, В.И.Фурса, Л.А.Еремина, В.И.Салыга

I. Общее описание

Статья является продолжением работы [I] и содержит результаты новой разработки - ППП КОД-2 ОС ЕС, выполненной по постановлению ГКНТ и принятой к эксплуатации межведомственной комиссией. Пакет КОД-2 представляет собой эффективный, доступный широкому пользователю программно-алгоритмический комплекс оперативного принятия решений посредством автоматизированной классификационной обработки данных (КОД) и построения структурно-аналитических моделей с помощью ЕС ЭВМ в АСУТП, САПР, задачах экспертного оценивания, группового выбора, научно-технического прогнозирования, технической и медицинской диагностики и др. Теоретическую основу пакета составляют структурно-аналитический метод распознавания образов с разнотипными признаками [2,3] *), структурно-аналитические модели и комбинаторные алгоритмы точной классификации данных в пространстве ранговых оценок [4,5].

Пакет КОД-2 предназначен для решения задач классификационной обработки данных с обучением и классификацией (К-задача), обработки данных с самообучением (Т-задача), классификационной обработки данных в пространстве ранговых оценок для принятия групповых решений и пакетной обработки анкет экспертного опроса (Э-задача). Математическая формулировка этих задач и алгоритмическое обеспечение их решения приведены в работах [I-6].

Пакет имеет простой входной язык директивного характера. Все прерывания в программах пакета замаскированы. В случае сбоя в ра-

*) Структурно-аналитический метод распознавания в идейной основе аналогичен подходу к анализу разнотипных данных, предложенному в работе [7]

боте пакета пользователю выдается соответствующее диагностическое сообщение (общее число сообщений 42) или повторяется запрос на консоль, если пакет работает в режиме диалога. Рабочая версия пакета содержит 136 программ (около 10 тысяч операторов языков ПЛ/I и ФОРТРАН), исследовательская версия - более 200 программ, включая развитую систему сервисных программ по обслуживанию пакета.

Функциональная схема пакета КОД-2 показана на рисунке. С системной точки зрения ППП КОД-2 есть задание ОС, состоящее из шагов P1-P4. Одиннадцать программных блоков (см. таблицу) составляют функциональное наполнение пакета и обеспечивают:

- ввод данных и настройку конфигурации пакета на решение Э-, Т-, К-задач (программа DISPOP1, главный диспетчер пакета);
- решение Э-задачи (подсистема "Эксперт", программа EXPERT2);
- решение Т-задачи (подсистема "Таксон", программа TAKSON3);
- решение К-задачи (подсистема "Классификатор", программа KLASS4).

Таблица

Блок	Назначение
INMAS	Ввод, контроль и корректировка данных с использованием режима диалога.
FORMTEX	Предобработка, печать и вычисление обобщенных характеристик таблицы эмпирических данных.
ANTIPOD	Поиск типологических и параметрических коалиций [4, 5], формирование правила классификации.
FORMKOL	Классификация, вычисление обобщенных характеристик и печать параметрических коалиций.
TAX	Таксономия объектов с количественными признаками методом адаптивных гистограмм, построение правила классификации.
MAKOL	Таксономия объектов малых таблиц эмпирических данных с количественными признаками методом делимой иерархической группировки, построение правила классификации.
TAKAP	Таксономия объектов с качественными и разнотипными признаками методом делимой иерархической группировки, построение правила классификации.
BLOCA	Генерация терминальных свойств-предикатов [1-3].
BLOCB	Формирование синтаксических образцов.
POSDER	Восстановление грамматики образов.
EXAMEN	Классификация объектов по найденному правилу классификации.

Функциональная схема пакета КОД-2.

Данные проблемной задачи классификационной обработки разделяются на входной язык и исходные данные. Входной язык состоит из набора директив и предложений. Общее число директив пакета - 5, предложений - 59. Предложения содержат закодированную информацию - постановку задачи классификационной обработки. Все предложения описаны в документе "Пакет прикладных программ КОД-2. Описание языка" [6] и хранятся в справочниках SLOWE, SLOWT, SLOWK (см. рисунок). Большинство предложений допускают умалчиваемые значения их параметров. Предложения могут следовать в произвольном порядке. Входной язык задачи классификационной обработки размещается на перфокартах. Исходные данные могут вводиться с перфокарт или из библиотеки на магнитных дисках.

Программа DISPOP1, в состав которой входит диспетчер-интерпретатор, осуществляет синтаксический и семантический контроль входного языка, формирует выходной унифицированный вектор параметров директивного запроса, содержащий как заданные, так и умалчиваемые значения параметров предложений, посылает вектор параметров соответствующей подсистеме.

В зависимости от выбранной задачи обработки данных главный диспетчер формирует определенный код возврата и передает управление требуемому шагу задания: P2, P3, P4. Исходные данные переписываются в набор данных &&DISTED, доступный всем шагам задания.

Подсистема "Эксперт" построена в виде оверлейной структуры из семи сегментов перекрытия и корневого сегмента EXPERT2, который является диспетчером Э-задачи.

Блоки подсистемы "Эксперт" реализуют следующие функции:

1. Блок INMAS - ввод параметров директивного запроса Э-задачи, ввод, логический и семантический контроль массива таблиц эмпирических данных и массива квалификационного состава экспертов, удаление ошибок и регламентированных пропусков из таблиц, корректировка данных и работа с генератором точек четырехмерного пространства ранговых оценок Π^4 в режиме диалога.

2. Блок FORMTEX - печать и предобработка таблиц эмпирических данных в пространстве ранговых оценок (вычисление рангов, кодов [4,5], поиск совпадающих объектов таблиц), вычисление и печать обобщенных статистических характеристик таблиц (коэффициентов конкордации, вариации и др.).

3. Блок ANTIPOD - поиск типологических коалиций объектов таблиц эмпирических данных в пространстве ранговых оценок, поиск

высокосогласованных пересекающихся параметрических коалиций, разделение пересечений и формирование иерархического правила классификации в виде списковой структуры, печать списка предикатов.

4. Блок **FORMKOL** - формирование и печать коалиций объектов обучающей таблицы данных, вычисление и печать обобщенных статистических характеристик коалиций, построение дендрограммы иерархической группировки коалиций обучающей таблицы. Если задана классификационная таблица данных, то для нее выполняются все перечисленные функции блока **FORMKOL** . Затем обучающая и классификационная таблица объединяются в одну, для которой повторно выполняются функции блока **FORMKOL**. Слияние таблиц позволяет проанализировать устойчивость правила классификации при появлении дополнительных данных (например, ответов новой группы экспертов), что представляет особый интерес в исследовательском режиме работы пакета КОД-2.

Подсистема "Эксперт" выполняет пакетную обработку произвольного числа таблиц эмпирических данных.

Подсистема "Таксон" построена в виде оверлейной структуры из шести сегментов перекрытия и корневого сегмента **TAKSON3**, который является диспетчером Т-задачи.

Блоки подсистемы "Таксон" реализуют следующие функции:

I. Блок **TAX** - поиск таксонов с заданными ограничениями, построение дендрограммы иерархической группировки таксонов (с заданным видом расстояния между таксонами) и графа смежности таксонов, описание структуры пространства признаков в виде бинарного дерева.

2. Блок **MAKOL** - таксономия объектов малых таблиц эмпирических данных с количественными признаками методами делимой или агломеративной иерархической группировки. При этом выполняются функции блока **TAX**. Если используется метод делимой иерархической группировки, то формируется правило классификации в виде бинарного дерева.

3. Блок **TAKAP** - поиск таксонов с заданной минимальной мощностью, построение правила классификации в виде бинарного дерева. При наличии разнотипных признаков выполняется отображение подпространства количественных признаков в качественные.

Подсистема "Классификатор" построена в виде оверлейной структуры из семи сегментов перекрытия и корневого сегмента **KLASS4** , который является диспетчером К-задачи.

Блоки подсистемы "Классификатор" реализуют следующие функции:

I. Блок **BLOCA** - генерация свойств-предикатов, безошибочно разделяющих классы, свойств-предикатов для признаков, измеренных в разнотипных шкалах, многомерных свойств-предикатов для двухклассовой задачи.

2. Блок **BLOCB** - формирование синтаксического образца, запись генерируемых свойств-предикатов в синтаксический образец, минимизация и сжатие образца.

3. Блок **POSDER** - построение правила классификации по синтаксическому образцу в виде бинарного дерева решений $G_B^{(2)}$, вершины которого есть свойства-предикаты, концевые вершины - номера классов.

4. Блок **EXAMEN** - ввод правила классификации из набора данных **KODGRAM** и таблиц данных для экзамена из набора данных **&&TEDE** (куда они были записаны диспетчером **KLASS4**), организация процедуры классификации объектов.

Рассмотрим примеры решения задач с помощью ППП КОД-2. В данной статье символ "|" условно является разделителем перфокарт.

2. Примеры решения Э-задачи

ПРИМЕР I. Входной язык имеет вид:

```
'ЭКСПЕРТ'|' Э:ИСТГЕН',  2 | 'Э:ИСТТЭД' , 3 | 'Э:КЛАССТЭД' , I
'Э:ПЕЧАТЬ' , 2 | 'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.
```

Запускается генератор точек четырехмерного пространства ранговых оценок $П^4$, формирующий ранговые оценки точек обучающей и классификационной таблиц данных. Номера точек оперативно вводятся с дисплея в режиме диалога. Размеры таблицы: число строк $M < 1000$, число столбцов $N = 4$. Выполняется предобработка таблицы в $П^4$, осуществляется поиск групп совпадающих точек и формирование обобщенных статистических характеристик последней. Результаты предобработки выводятся на печать по режиму полной выдачи.

По точкам обучающей таблицы в пространстве ПРО-4 алгоритм ANTIPOD [5] находит пересекающиеся однородные высокосогласованные параметрические коалиции, а алгоритм КРОТ [5], разделяя пересечения, формирует минимальный набор непересекающихся коалиций и выдает иерархическое правило классификации ранговых данных в виде списковой структуры (бинарного дерева), упорядоченное по убыванию мощности коалиций.

На печать выводятся коалиции и их обобщенные статистические характеристики. Строится дендрограмма иерархической группировки коалиций по заданной (явно или по умолчанию) мере близости. После этого выполняется классификация точек классификационной таблицы данных, результаты которой выводятся на печать. Отмечаются точки, не попавшие в коалиции.

В заключение, для того чтобы установить, насколько точки классификационной таблицы данных изменяют структуру коалиций обучающей таблицы, обе сливаются в одну, которая в результате "прогона" по найденному правилу классификации распадается на коалиции, для которых также вычисляются обобщенные статистические характеристики и печатается дендрограмма иерархической группировки. Выводимая информация в листинге структурирована, снабжена комментариями и наглядна для восприятия и анализа.

ПРИМЕР 2. Входной язык задачи:

```
'ЭКСПЕРТ' | 'Э:ИСТТЭД', 2 | 'Э:НАЧТЭД', 2 | 'Э:КОНТЭД', 6 |
'Э:ШКАЛАТЭД', 3 | 'Э:МОЩНОСТЬ', 5 | 'Э:МЕРАДГ', 2 | 'КОНЕЦ' |
'КОНЕЦ ДИРЕКТИВ'.
```

Используется режим таксономии данных группового выбора. Таблицы эмпирических данных расположены в библиотечном наборе данных BIMCOD на магнитных дисках. Все таблицы имеют количественные признаки. Входной язык предписывает пакетную обработку массивов: от таблицы (ТЭД) № 2 от № 6. Задана минимальная мощность коалиций - 5 точек. Размеры массивов определяются при их чтении. Дендрограмма иерархической группировки строится по мере 2 (минимальное локальное расстояние между коалициями). По умолчанию установлена сокращенная выдача на печать.

ПРИМЕР 3. Входной язык задачи:

```
'ЭКСПЕРТ' | 'Э:ЧИСЛОТЭД', I | 'Э:ДИАЛОГ', I | 'Э:ВЕСТЭД', 2 |
'Э:КЛАССТЭД', I | 'Э:КАЧЕСТВКЛАСС', I | 'Э:ПЕЧАТЬТЭД', I |
'Э:ПРОПУСКТЭД', 999 | 'Э:МАССИВКС', I | 'Э:ЧИСЛОКГ', 5 |
'Э:МОЩНОСТЬ', 5 | 'Э:ПЕЧАТЬ', 2 | 'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.
```

Входной поток содержит следующие массивы: квалификационный состав экспертов обучающей таблицы, весовые коэффициенты столбцов признаков таблицы, обучающую таблицу, массив квалификационного состава экспертов классификационной таблицы. При вводе массивов и наличии в них ошибок, элементы одномерных массивов корректируются в режиме диалога. Все массивы печатаются в исходном виде в порядке их размещения во входном потоке. Массивы таблиц содер -

жат регламентированные пропуски, обозначенные числом 999. При обработке таблиц проводится качественная классификация и формируются типологические коалиции. Для параметрических коалиций строится дендрограмма иерархической группировки по мере I (статистическое расстояние). Результаты классификации печатаются в полной форме.

3. Примеры решения Т-задачи

ПРИМЕР 4. Входной язык задачи:

```
'ТАКСОН' | 'Т:МОЩНТАКС', 10 |'Т:ТИПДГ', 1 | 'Т:ПЕЧАТЬТАКС', 0 |
'Т:РЕЖТАКС', 2 | 'Т:ВИДТЭД', 2 |'Т:ПЕЧАТЬ', 2 | 'Т:ТЭД', 98,2 |
'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.
```

Вводится таблица данных с количественными признаками (98 объектов, 2 признака), элементы которой расположены по столбцам. При помощи алгоритма адаптивных гистограмм находятся таксоны. На найденных таксонах строится дендрограмма иерархической группировки, на которой определяются таксоны с мощностью, не меньшей заданного порога, равного 10. В качестве меры близости таксонов используется статистическое расстояние. Строится граф смежных таксонов. Генерируется описание пространства признаков в виде бинарного дерева, представленного списковой формой. Определяются области таксонов и граничные области, разделяющие таксоны. Выводится список номеров точек таксонов.

ПРИМЕР 5. Входной язык задачи имеет вид:

```
'ТАКСОН' |'Т:ТИПДГ', 1 | 'Т:МОЩНТАКС', 20 | 'Т:ВИДТЭД', 2|
'Т:ТИППРИЗ', 2 | 'Т:ТИПДГК', 4 | 'Т:КАЧПР', 6|'Т:МЕРАК', 2|
'Т:ТЭД', 80 , 8 | 'Т:ПЕЧАТЬТАКС', 0 | 'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.
```

Вводится таблица с разнотипными признаками (число объектов - 80, число признаков - 8). Шесть признаков - качественные. Таблица данных расположена на перфокартах по столбцам.

В подпространстве количественных признаков определяются таксоны с минимальной мощностью в 20 точек. Для нахождения таксонов используется алгоритм ДИГ. Строится дендрограмма иерархической группировки найденных таксонов. Для оценки близости таксонов используется статистическое расстояние. Строится правило классификации таксонов в пространстве количественных признаков и граф смежных таксонов. Посредством построенной дендрограммы количественные признаки отображаются в качественные. В пространстве качественных признаков определяются таксоны с минимальной мощностью в 10 точек. Строится дендрограмма иерархической группировки таксо-

нов с мерой близости, использующей минимальное локальное расстояние. Генерируется правило классификации найденных таксонов. Выводится список номеров точек таксонов.

4. Примеры решения К-задачи

ПРИМЕР 6. Входной язык имеет вид:

'КЛАССИФИКАТОР' | 'К:ТЭД' 50, 7, I0 | 'К:ПОРЯДОК' II |
'К:ЗАПИСЬК' I | 'К:СПМИН' I | 'К:СПМАХ' I | 'К:ГИСТ' I |
'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.

Вводится таблица с разнотипными признаками (тип шкалы измерения задается в массиве шкал). Задан базовый набор свойств-предикатов СПМАХ, СПМИН, СПГИСТ, СПЛДФI, BAIS - последние два - по умолчанию. Число объектов - 50. Число признаков - 7. Число классов - I0.

Строится правило классификации, на печать выводятся:

- таблица для обучения;
- параметры свойств-предикатов (таблицы);
- дерево решений;
- обобщенные характеристики правила классификации. Перед окончанием выполнения задачи правило классификации записывается для хранения в набор данных KODGRAM.

ПРИМЕР 7. Входной язык задачи:

'КЛАССИФИКАТОР' | 'К:ТЭД' I50, I0, 2 | 'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.

Вводится таблица с разнотипными признаками. Число объектов - I50, число признаков - I0, число классов - 2. Строится правило классификации, выводится информация, описанная в примере 6.

ПРИМЕР 8. Входной язык имеет вид:

'КЛАССИФИКАТОР' | 'К:ТЭДКЗ' I00, 7 | 'КОНЕЦ' | 'КОНЕЦ ДИРЕКТИВ'.

Решается задача экзамена - классификации объектов экзаменационной выборки по полученному ранее правилу классификации.

Вводится таблица для экзамена, содержащая I00 объектов и 7 признаков. Правило классификации вводится из набора данных KODGRAM. Проводится классификация и выводится следующая информация:

- таблица для экзамена;
- таблица параметров свойств-предикатов;
- дерево решений;
- распределение объектов выборки по классам.

5. Заключение

Пакет КОД-2 функционирует под управлением ОС в пакетном и диалоговом режимах (пакетный режим - основной, диалоговый - исследовательский) и не требует монополии на ресурсы ЭВМ. Для нормальной работы пакета требуется 256 Кбайт оперативной памяти. Время решения одной задачи классификации данных зависит от размера таблиц данных; среднее время составляет 2-20 минут. В качестве инструментальной ЭВМ использована ЭВМ ЕС-1033 с ОС версии 6.1. Документация на пакет выполнена в соответствии с требованиями ГОСТ ЕСПД. Пакет сдан в ГосФАП [6], успешно опробован на многочисленных тестовых и практических задачах и внедрен с реальным экономическим эффектом на ряде промышленных предприятий, в том числе на Новолипецком металлургическом заводе. С помощью ППП КОД-2 решены задачи распознавания производственных ситуаций при управлении технологическими процессами металлургических объектов, экспертной оценки параметров человеко-машинных комплексов при их разработке и эксплуатации, экспертной оценки деятельностии госнадзора за внедрением и соблюдением стандартов, ТУ и качеством промышленной продукции в системе Госстандарта, комплекс задач классификации селекционных признаков зернобобовых культур (по Продовольственной программе СССР), отдельные задачи технической и медицинской диагностики.

Литература

1. Пакет прикладных программ классификационной обработки данных ППП КОД-2 ДОС ЕС/И.Б.Сиродка, Н.Г.Голубь, Е.М.Крылов,и др. -В кн.: Математические методы анализа динамических систем. Харьков, 1979, вып. 3, с. 111-128.

2. СИРОДЖА И.Б. Структурно-аналитический метод распознавания образов с разнотипными признаками. -В кн.: Математические методы анализа динамических систем. Харьков, 1981, вып. 5, с. 91-107.

3. СИРОДЖА И.Б. Системный синтез структурно-аналитических алгоритмов распознавания для автоматизации классификационной обработки данных (КОД). -В кн.: Математические методы анализа динамических систем. Харьков, 1978, вып. 2, с. 79-102.

4. ВОЛКОВИЦКИЙ К.Е., КРЫЛОВ Е.М., СИРОДЖА И.Б. Анализ пространства ранговых оценок и синтез структурно-аналитических моделей для автоматизации принятия экспертных решений. -Киев, 1981. - 59 с. -(Препринт/Ин-т кибернетики АН УССР: 81-43).

5. КРЫЛОВ Е.М. Математическое и программное обеспечение классификационной обработки данных в пространстве ранговых оценок для автоматизации принятия решений. -В кн.: Математические методы анализа динамических систем. Харьков, 1983, вып. 7, с. 118-124.

6. Пакет прикладных программ КОД-2 /И.Б.Сиродка, Е.М.Крылов, В.А.Дискант, В.И.Фурса, Л.А.Еремина. - Харьков, 1982. -446 с.-Рукопись представлена Харьковским авиационным институтом. Деп.в РФАП СКТБ ПО ИК АН УССР 29 дек. 1982, № 6029.

7. ЛБОВ Г.С., КОТЮКОВ В.И., МАНОХИН А.Н. Об одном алгоритме распознавания в пространстве разнотипных признаков. -Вычислительные системы. Вып.55. Новосибирск,1973,с.98-107.

Поступила в ред.-изд.отд.
5 апреля 1984 года