ВОПРОСЫ ОБРАБОТКИ ИНФОРМАЦИИ ПРИ ПРОЕКТИРОВАНИИ СИСТЕМ
(Вычислительные системы)
1977 год Выпуск 69

УДК 007:681.3.068:519.3

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ ОБРАБОТКИ ТАБЛИЦ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ОТЭКС-I

Н.Г.Загоруйко, Г.С.Лбов, Ю.П.Маиаров

Одним из направлений развития проблемно-ориентированного программного обеспечения является создание пакетов прикладных программ. Использование пакета существенно снижает требования к квалификации программиста-пользователя, что дает возможность уменьшить время между окончанием научных разработок и внедрением их в практику.

В Институте математики СО АН СССР и Новосибирском государственном университете создан пакет прикладных программ ОТЭКС-I для многоцелевой обработки экспериментальных данных, представленных в табличном виде. ОТЭКС-I является методоориентированным пакетом, предназначенным для решения задач из областей медицины, геологии, экономики, социологии. Он содержит набор частично взаимосвязанных программ и имеет простую структуру с запланированным перекрытием. Понятия, используемые для характеристики пакетов прикладных программ, можно найти в работах [1,2]. Программы, входящие в пакет, оформлены в соответствии с основными положениями инструкции Государственного фонда алгоритмов и программ.

Сам пакет оформлен согласно ОСТу 25-231-74 "Автоматизированные системы управления. Система пакетов прикладных программ. Техническая документация. Виды, комплектность, содержание". Полное описание пакета ОТЭКС-I и распечатки программ приводятся в работе [3]. Пакет реализован для ЭВМ "Минск-32".

ОТЭКС-I рассчитан на широкого пользователя, для которого выбор наиболее подходящей программы из пакета в зависимости от решаемой задачи является затруднительным. Для такого выбора требуются достаточно полные знания о каждом реализованном методе ре-

мения и условиях его применения. Предлагаемый вариант пакета обеспечивает автоматический выбор режима вычислений по достаточно простому запросу.

Решаемая задача характеризуется восемью параметрами. Набор значений параметров образует то или иное управляющее слово (запрос).

Для характеристики прикладных задач выбраны следующие параметры:

Тип решаемой задачи. Программы пакета осуществляют решение следующих задач обработки таблиц экспериментальных данных:

- таксономия (автоматическая группировка объектов и признаков);
- выбор наиболее информативной подсистемы признаков из исходной системы;
- построение решающего правила для распознавания, в том числе последовательной процедуры принятия решения;
- планирование и обработка экспериментов при поиске приближенного значения глобального экстремума функции;
- восстановление многоэкстремальной функции;
- заполнение пропусков в таблицах экспериментальных данных.

Тип шкалы. Необходимость введения этого параметра вызвана существованием большого класса эмпирических таблиц в областях медицины, геологии, экономики и социологии, характерной особенностью которых является наличие разнотипных признаков. Методы обработки должны учитывать типы шкал, в которых замерены признаки, и быть инвариантными по отношению к допустимым преобразованиям этих шкал [4]. Этот параметр принимает различные значения, если исходные признаки в таблице замерены: а) в шкалах интервалов, отношений и абсолютной шкале, б) в шкалах различных типов.

Число признаков (n). Этот параметр определяет выбор различных программ обработки для случаев: а) $n \leq 20$; б) $20 < n \leq 100$; в) $n > 100$.

Число реализаций (N). Этот параметр принимает значение, соответствующее числу реализаций (объектов) в обучающей выборке, и определяет выбор различных программ для случаев: а) $N \leq 100$, б) $100 < N \leq 200$, в) $N > 200$.

Наличие пропусков в таблице. Параметр принимает одно из двух значений в зависимости от того, присутствуют или нет пропуски в экспериментальной таблице, и также влияет на выбор той или иной программы.

Характеристика зависимости признаков. Выбор одного из двух значений этого параметра зависит от того, принимается или не принимается заранее пользователем гипотеза о независимости признаков.

Ограничение на форму таксона. Этот параметр имеет смысл только для задач таксономии и принимает различные значения в зависимости от того, какие ограничения накладываются на форму таксона в задачах таксономии: а) сферические таксоны (алгоритм "Фораль" [5]), б) таксоны произвольной формы (алгоритм "Краб" [6]);

Машинный носитель. Этот параметр принимает два различных значения в зависимости от того, перфокарты или магнитная лента являются машинным носителем массива экспериментальных данных.

Ниже приводятся краткие аннотации к алгоритмам, реализованным в пакете прикладных программ ОТЭКС-I .

Алгоритм "Фораль" [5] предназначен для решения задачи таксономии в случае количественных признаков*). В качестве формы таксонов выбраны гиперсферы. Алгоритм позволяет найти заданное число гиперсфер минимального радиуса, дающих разбиение данного множества реализаций, или минимальное число гиперсфер при заданном их радиусе.

Алгоритм "Краб" [6] в отличие от алгоритма "Фораль" строит таксоны произвольной формы. Для этого сначала все точки пространства признаков соединяются кратчайшим незамкнутым путем. Далее границы между таксонами устанавливаются на некоторых выбранных отрезках кратчайшего незамкнутого пути. Алгоритм "Краб" позволяет найти наилучший вариант выбора границ с точки зрения некоторого формализованного критерия качества таксономии, учитывающего близость точек внутри таксона, удаленность таксонов друг от друга, равномерность распределения числа точек в таксоне по таксонам и т.д. Ограничения на форму таксонов при этом не используются.

*) Количественными признаками здесь названы признаки, замеренные в абсолютной шкале и шкале отношений и интервалов, качественными признаками – в шкале порядка; классификационными – в шкале наименований [4].

Алгоритм НТПП [7] - алгоритм направленного таксономического поиска информативных подсистем признаков - производит таксономию признаков (в качестве меры близости используется коэффициент корреляции между признаками).Из каждого таксона выбирается по одному "типичному" (ближайшему к среднему) представителю. Далее,формируются различные подсистемы из "типичных" признаков. Наилучшей подсистемой признаков считается та, для которой получено минимальное число эталонов по методу дробящихся эталонов (ДРЭТ [8]). В данном случае предполагается, что все признаки количественные.

Алгоритм СПА-I [9] также предназначен для выбора наиболее информативной подсистемы признаков из некоторой исходной системы количественных признаков. Задача выбора признаков ставится как экстремальная задача на единичном гиперкубе. В качестве множества вариантов (вершин гиперкуба) выступает множество всевозможных сочетаний признаков из исходной системы. Для каждого такого варианта на основе обучающей выборки определяется квадратичное решающее правило. Качество варианта (подпространства) оценивается по числу ошибок на обучающей выборке. Алгоритм СПА-I реализует поиск наилучшего варианта путем сочетания метода Монте-Карло с "поощрением" и "наказанием" признаков во время поиска.

Алгоритм ДРЭТ [8]. Решающее правило у алгоритма реализуется в виде набора гиперсфер. Для каждого из образов определяется сфера минимального радиуса, включающая все реализации обучающей выборки. Если сферы пересекаются,то для объектов,попавших в пересечение, вновь строятся сферы меньших радиусов и т.д. Процесс построения сфер прекращается, если в возможных пересечениях нет объектов из разных образов. Предполагается, что все признаки количественные. Контрольная реализация относится к тому образу, в сферу которого она попадает.

Алгоритм ТРФ [8] в качестве решающей функции использует таксономические критерии (те же, что в "Крабе"). Предполагая, что контрольная реализация относится к данному образу, строим кратчайший незамкнутый путь для рассматриваемого образа с учетом данной реализации. Строятся участки кратчайшего незамкнутого пути, соединяющего между собой все образы.Контрольная реализация относится к тому образу, присоединение к которому данной реализации дает наибольшее значение критерия качества таксономии. Предполагается, что все признаки количественные.

Алгоритмы "Коралл" и ТЭМПР [10,11] предназначены для выбора наиболее информативной подсистемы признаков, замеренных в разных шкалах. Результатом работы обоих алгоритмов является набор информативных логических высказываний для каждого образа.Под логическим высказыванием понимается конъюнкция значений и интервалов признаков. Высказывание для данного образа считается информативным, если оно часто выполняется на этом образе и редко на других. Признаки, не вошедшие ни в одно из выбранных высказываний,исключаются из исходной системы как неинформативные.Допускаются пропуски в эмпирических таблицах.

Алгоритм ТЭМПР выделяет практически все информативные высказывания. Однако программа работает при сравнительно небольшом числе признаков исходной системы ($n \leq 20$). Алгоритм "Коралл" реализует некоторую направленную процедуру перебора высказываний, при которой могут не выделиться некоторые информативные высказывания. Однако его использование позволяет решить задачу при большом числе признаков ($n \approx 100$).

Алгоритм ДРП [3,11] предназначен для формирования и использования решающего правила на основе списка выделенных информативных логических высказываний. Допускаются пропуски в таблице данных.

Алгоритм "Ранес" [3] предназначен для выбора информативной подсистемы признаков и принятия решения для распознавания образов в случае независимых признаков, замеренных в разных шкалах. Допускаются пропуски замеров в таблице экспериментальных данных.

Алгоритм ИРС-I [12] реализует последовательную процедуру принятия решения о принадлежности к образу, предназначенную для минимизации числа измерений признаков у распознаваемого объекта. Считается, что обучение уже проведено и выбрана наиболее информативная система признаков. В общем случае для каждого распознаваемого объекта существует свое (достаточное для принятия решения) подмножество признаков, которое желательно выбрать.

Последовательная процедура принятия решения реализована в предположении, что образы описываются нормальными распределениями с неравными матрицами ковариаций в пространстве количественных признаков.

Алгоритм СПА-2 [13] реализует процедуру адаптивного поиска приближенного глобального экстремума функции от переменных, замеренных в шкале наименований. Адаптивный поиск предусматривает раз-

биение общего числа экспериментов (вычислений значений функции) на ряд групп. Общее число экспериментов устанавливается заранее. После проведения каждой группы экспериментов исключаются "худшие" значения каждой из переменной. Тем самым после каждой группы экспериментов сокращается допустимая область значений переменных.

Алгоритм LL [I4] реализует процедуру адаптивного поиска приближенного глобального экстремума функции, удовлетворяющей условию Липшица. Причем по мере проведения поиска осуществляется оценка константы Липшица. Допустимая область переменных представляет собой гиперпараллелепипед. Общее число экспериментов устанавливается перед поиском. Признаки количественные.

Алгоритм "Прогноз" [3] решает задачу восстановления много - экстремальной функции по экспериментальным данным. Для решения задачи используется метод потенциальных функций с оптимизацией их параметров и с одновременным выбором наиболее существенной подсистемы переменных. Предполагается, что все признаки количественные.

Алгоритм "Пинч" [3] предназначен для выбора наиболее информативных логических высказываний с целью восстановления функции (предсказания значения количественного признака с точностью до заранее фиксированных интервалов) в случае признаков, замеренных в разных шкалах. Алгоритм допускает пропуски в эмпирических таблицах.

Алгоритм "ZET-75" [I5] предназначен для заполнения пропусков в эмпирических таблицах. Идея алгоритма состоит в следующем. Фиксируется строка и столбец, на пересечении которых находится пропуск. Затем выделяются "к о м п е т е н т н ы е" (похожие) столбцы и строки, по которым (в предположении линейной зависимости между строками) делается предсказание значения пропущенного элемента. Признаки количественные.

В разработке отдельных алгоритмов и программ, входящих в состав ОТЭКС-I , принимали участие Вербкина В.П., Ёлкина В.Н., Загоруйко Н.Г., Лбов Г.С., Лебедев В.Г., Машаров Ю.П., Тимеркаев В.С.

В настоящее время разрабатывается второй вариант пакета прикладных программ ОТЭКС . Эта разработка обусловлена необходимостью создания пакета с более гибкой внутренней структурой, с большим сервисом для пользователя и с большим набором программ.

Во втором варианте пакета предусматриваются возможности:

- диалогового режима на всех этапах решения задачи, что даст возможность пользователю влиять на процесс вычислений;
- диалога с информатором, выдающим необходимую справочную информацию о пакете на пультовую пишущую машинку и АЦПУ;
- работы с пакетом как с библиотекой модулей, так и с автоматическим выбором режима вычислений (в зависимости от квалификации пользователя);
- генерирования логической структуры комплекса с использованием метода таблиц решений;
- динамического распределения ресурсов ЭВМ в процессе работы пакета.

Второй вариант пакета ОТЭКС реализуется на ЭВМ "Минск-32" и ЕС ЭВМ и рассчитан на пополнение программами различных авторов и организаций, представленными в унифицированном виде.

Кроме указанного сервиса, создание в будущем большого пакета обработки таблиц экспериментальных данных, объединяющего программы различных научных коллективов, даст возможность сравнивать различные алгоритмы по их эффективности.

Большое число существующих в настоящее время алгоритмов и отсутствие критерия их эффективности не позволяют пользователю выбрать наилучший из них. Кроме того, отсутствие возможности такого сравнения приводит к созданию новых алгоритмов, эффективность которых может быть не выше уже существующих. Сравнение алгоритмов вне рамок пакета связано с известными теоретическими и практическими трудностями. По нашему представлению, сравнение алгоритмов может происходить следующим образом.

Указанный набор параметров (тип задачи, тип признаков, размерность пространства, объем выборки и т.п.) выделяет некоторое подмножество прикладных задач и набор конкурирующих алгоритмов для их решения. В результате решения каждой конкретной задачи пользователь указывает, какому из использованных алгоритмов отдать предпочтение, исходя из критерия качества (например, числа ошибок на экзамене), времени и памяти ЭВМ, которые требовались при обучении и контроле, интерпретируемости правила принятия решения, сложности технической реализации этого правила и т.д.

Необходимо обратить особое внимание на интерпретируемость решающего правила. Именно это свойство может расширить возможности диалогового режима. Кроме того, в тех задачах, когда качество

решения определяется научной ценностью полученного решающего правила, интерпретируемость его может служить основным критерием оценки алгоритма.

Эксплуатация пакета многими пользователями приведет к тому, что со временем появится возможность установления предпочтения среди множества конкурирующих алгоритмов по результатам их многократного одновременного использования и указания наиболее эффективных условий применения тех или иных алгоритмов.

Литература

I. АЛФЕРОВ З.В., ДИХАЧЕВА Г.Н., ШУРАКОВ В.В. Математическое обеспечение ЭВМ. М., "Статистика", I974.

2. ФАТЕЕВ А.Е., РОЙТМАН А.И., ФАТЕЕВА Т.П.Прикладные программы в системе математического обеспечения ЕС ЭВМ. М., "Статистика", I976.

3. Пакет прикладных программ для обработки таблиц экспериментальных данных "ОТЭКС-I", Новосибирск, НГУ, I977.

4. СУППЕС П., ЗИНЕС Дж. Основы теории измерений. -В кн.: Психологические измерения . М., "Мир", I967, с. 9-II0.

5. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г. Об алфавите объектов распознавания. -В кн.: Вычислительные системы. Вып. 22. Новосибирск,I966, с. 59-76.

6. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений.-В кн.: Вычислительные системы. Вып. 36. Новосибирск,I969,с.29-46.

7. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г., ТИМЕРКАЕВ В.С. Алгоритм направленного таксономического поиска информативных подсистем признаков (НТПП). -В кн.: Вычислительные системы. Вып. 59. Новосибирск, I974, с. 49-70.

8. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. М., "Сов.радио", I972.

9. ЛБОВ Г.С. Выбор эффективной системы зависимых признаков. -В кн.: Вычислительные системы. Вып. I9. Новосибирск, I965, с.2I-34.

I0. ЛБОВ Г.С., КОТЮКОВ В.И., МАНОХИН А.Н. Об одном алгоритме распознавания в пространстве разнотипных признаков. -В кн.: Вычислительные системы. Вып. 55. Новосибирск, I973, с. I08-II0.

II. ЛБОВ Г.С., КОТЮКОВ В.И., МАШАРОВ Ю.П. Метод поиска логических закономерностей на эмпирических таблицах. -В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы, вып. 67.) Новосибирск, I976, с. 29-4I.

I2. ЛБОВ Г.С., КОТЮКОВ В.И. Последовательная процедура распознавания для случая коррелированных измерений. -В кн.: Вычислительные системы. Вып. 6I. Новосибирск, I975, с. 37-42.

I3. ЛБОВ Г.С. Об одном алгоритме поиска экстремума функции от переменных, замеренных в шкале наименований. -В кн.: Вычислительные системы. Вып. 44. Новосибирск, I97I, с. I3-22.

I4. ЛБОВ Г.С., ТРУНОВ А.А. Об одном алгоритме поиска глобального экстремума функции. -В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы, вып. 67.) Новосибирск, I976, с. 29-4I.

I5. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ТИМЕРКАЕВ В.С. Алгоритм "ZET-75" заполнения пробелов в эмпирических таблицах. -В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы, вып. 67.) Новосибирск, I976, с. 3-28.

Поступила в ред.-изд.отд.
2I января I977 года