**Анализ машиночитаемых документов компьютерными средствами**

Е.В. Злобин, И.В. Попенков

Расширение компьютерного инструментария историков - задача не только важная, но и достаточно сложная в силу своей специфичности, трудности наработки программного обеспечения для "чистого" обществоведа. Система KLEIO, программы проф. Яна Олдерволла представляются скорее исключением, подтверждающим общее правило. Предлагаемая работа претендует на попытку разработки собственного программного обеспечения, в чем-то дополняющего имеющиеся пакеты математической статистики, в то же время, ввиду специфики своей разработки, ориентированного на ввод и анализ качественных признаков, измеренных в шкале наименований, которая в абсолютном большинстве случаев используется историками.

Другим побудительным мотивом для данного исследования явился кризис исторической информатики, о котором пишет П. Доорн в своих известных тезисах. Он в какой-то степени должен стимулировать интерес к методической стороне процесса, а именно, к методам исследования полученных в ходе грандиозных "набивок" баз данных. Нами описывается одна из систем для анализа больших массивов структурированных данных смешанной - числовой и нечисловой - природы.

Весьма часто при изучении исторических явлений или процессов приходится иметь дело с задачами классификации объектов по одному количественному признаку - числовому - (измеренному в шкале отношений), который является выходным, и по нескольким признакам, измеренным в шкале наименований (входным).

Задача классификации - минимизировать дисперсию выходного признака по каждому входному. Решается она в ряд этапов. На первом этапе первоначальная матрица данных разбивается на группы. При этом по выходному признаку и одному из входных выбирается такое разбиение, при котором сумма внутригрупповых дисперсий минимальна. Объекты могут попадать в разные группы с одним и тем же значением входного признака. Затем идет анализ распределения внутри каждой из групп. Тем самым число входных признаков уменьшается на единицу. Процесс этот идет до тех пор, пока изменение суммы внутри группового признака становится минимальным (менее заданного порога). В результате получаются однородные по всей совокупности входных признаков и по их отношению к выходному признаку группы.

Типично исследовательской задачей такого рода является изучение влияния возраста на формирование тех или иных групповых качественных характеристик исторических личностей. Подобного рода анализ проводился одним из авторов при выявлении внутренней структуры высшего выборного органа КПСС - ее Центрального Комитета и Политбюро последнего, "предсмертного" состава. Уже тогда было ясно, что реальный математический и программный аппарат для такого рода задач не наработан.

Особенно ценным представляется использование данного подхода при анализе просопографических баз данных, которые зачастую с трудом поддаются формализации и количественной оценке, но в которых возрастная графа наличествует обязательно. Помимо чисто возрастной графы в просопографических базах присутствуют различные поля типа даты (времени) получения различных должностей (образования, наград и пр.), которые позволяют вычислить срок пребывания в той или иной категории. Данные количественные показатели уже позволяют применить описанные выше подходы.

Предлагаемый подход к классификации обьектов, описанных качественными признаками, может быть полезен и потому, что наиболее часто применяемые методы кластерного анализа, в общем случае, пригодны только для шкал отношений. Для других шкал возможно использование т.н. информационного подхода, но это не всегда является удобным, и приводит, по видимому, к некоторому огрублению полученнного результата[1].

Авторами доработана система[2], позволяющая в масштабе реального времени решать задачи классификации для 5 и более (опробована модель на 15) тыс. обьектов. Программа реализована на 32-битном FORTRANe, в соответствии со стандартом FORTRAN 90 [3]. При использовании ее процессор переводится в защищенный режим работы, тем самым позволяя использовать реально всю физическую память, установленную на машине. Имеется возможность компиляции программы специально для использования под Windows в расширенном режиме, которая снимает все ограничения на объемы используемых массивов.

Состав системы приведен в приложении 1. Входными для программы являются файлы типа ASCII, которые реализуются исходя из имеющихся баз данных, и один рассчитываемый файл прямого доступа, а также задаваемый исследователем критерий изменения дисперсии в группе (от 0,01 до 0,5). Выходным - обычный текстовый файл с подробный распечаткой результатов расчета.

Структурная схема работы программы представлена на Рис. 1. Она тривиальна и подчеркивает особенность системы - ввод наименований нечисловых признаков (т.н. словарей значений).

Рис. 1

Подсистема ведения архива также решает стандартные для любой СУБД задачи. Особенностью ее является наличие дружественного интерфейса и ориентированность на подготовку данных для собственно расчетных задач (см. Рис. 2).

Рис. 2.

Специально выделенная подсистема работы со словарями также выполняет тривиальные задачи СУБД с одним ограничением - ориентированностью на ввод наименований признаков, ограниченных одной строкой не более 40 символов. Данное ограничение вызвано реальными размерами экрана машины и большим расходом оперативной памяти на массив наименований значений. Функции ее раскрыты на Рис. 3.

Рис. 3

Подсистема классификации проводит классификацию на основе информации, содержащейся в записях и словарях базы данных. Классификация состоит в разбиении заданного множества записей на непересекающиеся по входным признакам классы так, чтобы суммарная по всем классам дисперсия выходного признака была минимальна.

Выходной и входные признаки, характеризующие объекты исследования, берутся из записей базы (общие для всей совокупности ), причем выходной признак обязательно числовой (например, возраст, объем доходов, численность работающих, населения и т.п.), а входные признаки могут быть и качественными, но задаются символьными кодами (до 15 символов).

Диалог пользователя с системой строится с вводом следующей информации: - название базы данных; - определение совокупности записей из архива для проведения классификации; - выходной (числовой ) признак, по которому проводится классификация; - набор входных признаков, с помощью которых описываются группы; - параметр разбиения, определяющий число групп.

Эксплуатация данного программного продукта максимально проста. При работе с программой для выбора в системе меню используются клавиши дополнительной клавиатуры, а для подтверждения выбора клавиша . Для выхода и продолжения действия в большинстве пунктов используется клавиша . Перечень используемых клавиш постоянно приводится в нижней строчке экрана.

Описываемая программа предъявляет жесткие требования к "железу" компьютера и предназначена для использования на IBM-совместимых ПЭВМ, имеющих процессор с разрядностью не ниже 32 (то есть, 386 и выше) и требует для работы не менее 4 мБ общей ОЗУ и не менее 600 кБ в нижней памяти (conventional). Данное ограничение на память введено ввиду огромной размерности просчитываемых массивов (в данной версии число объектов 5000).

Для получения максимально возможного обьема памяти рекомендуется использовать операционную систему версии 6.2 и выше. Для освобождения "нижней" памяти необходимо драйверы устройств загружать в "верхнюю" память и в блоки UMB, туда же помещать и операционную систему. Точно также требуется и максимально возможное количество резидентных программ грузить "наверх". Все это достигается использованием драйверов HIMEM.SYS и EMM386.EXE, команд DEVICEHIGH и LH, реализуемых стандартным образом.

Драйвер EMM386 запускается с ключом NOEMS. В случае выдачи сообщения о нехватке памяти при запуске базы возможны следующие пути преодоления этой проблемы: - включение ключа NOVCPI в командную строку драйвера EMM386; - включение строки SET CLIPPER=E в файл AUTOEXEC.BAT.

В приложении N 2 предлагаются варианты конфигурации операционной системы, применительно к высказанным рекомендациям на компьютера с ОЗУ 4 мБ, включенной звуковой картой, CD-диском и прозрачным архиватором винчестера STACKER. В случае потребности в другой конфигурации ОС на ПЭВМ, имеет смысл реализовать данные файлы с меню выбора при загрузке.

Опишем кратко расчетные программы. Программа VVPR формирует вектор-запись в файл прямого доступа OBPR. Содержательно эта запись представляет собой значения очередного признака для всех выбранных объектов.

Обращение к программе VVPR происходит в цикле (по числу признаков) из базовой программы при обработке выбранных признаков

Вход:

|  |  |  |
| --- | --- | --- |
| имя файла | размер | содержимое |
| PARAM.TXT | 1 запись-4 поля | nob - число выбранных объектовi1 - номер текущего признакаnnopr - название выходного признакаeps - точность |
| PROB.TXT | 1 запись - nob полей | значения текущего признака для всех выбранных объектов |

Выход:

Файл прямого доступа OBPR (матрица объект-признак) строка - признак, столбец - объект, размерность - число выбранных объектов, умноженное на число выбранных признаков, первая строка - значения выходного признака для всех объектов, остальные строки - значения соответствующих входных признаков для всех выбранных объектов (в ходе дальнейшей работы стирается).

Программа VVPR за один проход формирует одну запись типа строка - признак и готовит данные для работы программы RASH - собственно расчетную программу. У этой программы:

Вход:

Файлы OBPR, PARAM.TXT, NAPR, NAZPR.

Файлы NAPR и NAZPR формируются в программе базы на основе информации, содержащейся в словарях.

Выход:

Файл FCSG.DAT Этот файл представляет собой одну запись, состоящую из следующих элементов:

|  |  |
| --- | --- |
| 1 | Номер группы |
| 2 | Число объектов в данной группе=nobg |
| 3-2+nobg | Номера объектов в группе |
| 2+nobg+1 | Номер уровня, на котором произошло разбиение по этому признаку |
| 2+nobg+2 | Номер признака |
| 2+nobg+3 | Число значений данного признака |
| следующие jpr элементов | Номера значений признака |

Затем номер следующей группы, далее все аналогично. Номера уровней идут по убыванию, как только номер уровня=1, начинается следующая группа.

На входе в RASH также:

Файлы:

napr.txt - имена признаков, которые мы выбрали из словаря словарей

nazpr.txt - название значений признаков из соответствующего словаря, код словаря по каждому признаку

Берутся подряд все признаки, сначала числовой код, затем названия признака подряд все перечисляются.

Файл PARAM.TXT - 500\_3\_"выходной признак"\_0,200. Включает: Число записей (500), число входных признаков (3), название выходного признака, необходимую точность вычислений (0.2).

Файл BNAPR.DBF - следующие поля:

NSLOV, LSLOV - число записей в словаре, NAPRIZ - название словаря, NSL - имя словаря.

STRA.DBF - файл dbf, в который записываются выбранные параметры построчно.

На выходе системы формируется текстовый файл итоговых расчетов. Пример его приведен в Приложении 3. В данной версии этот файл затирается, но может быть легко восстановлен стандартными средствами (типа UNDELETE и пр.). В дальнейшем он может редактироваться любым текстовым редактором.

Описанная система будет использована при обсчете больших массивов информации, нарабатываемх в ходе реализации совместных проектов с Государственным архивом РФ.

Приложение 1. Состав системы.

Система помещается на 1 дискете 5,25' (1,2 мБ) или 3,5' (1,44 мБ) и включает следующие файлы:

|  |  |  |
| --- | --- | --- |
| VVPR.EXEDOS4GW.EXERASCH.EXE |  | Расчетные программы для классификации. |

|  |  |  |
| --- | --- | --- |
| ITOG.DBFKAT\_A.DBFKAT\_S.DBFPROSM.DBFSHAB\_A.DB\_SHAB\_S.DBFITG\_D.DBFSHAB\_AD.DBF |  | Файлы баз данных необходимых для нормального функционирования системы. |

DIAG\_DEM.EXE - Демонстрационная программа

RECLAMA.EXE - Рекламный ролик с музыкой

ITG\_D.SYS - Текстовый файл - пример результата полученного в ходе классификации (необходим для нормального функционирования демонстрационной программы)

BRED.BAT - Специфичный файл, не требующий запуска пользователем. Необходим для нормальной работы системы.

Приложение 2. Варианты написания файлов config.sys и autoexec.bat:

Содержимое файла config.sys

DEVICE=C:\DOS\HIMEM.SYS

DEVICEHIGH=C:\DOS\EMM386.EXE NOEMS NOVCPI /V

DEVICE=C:\STACKER\DPMS.EXE

DEVICEHIGH=C:\STACKER\STACHIGH.SYS

LASTDRIVE=H DOS=HIGH,UMB

rem Загрузка DOS наверх

FILES=100

DEVICEHIGH=C:\DOS\SETFNT.SYS

COUNTRY=07,,C:\DOS\COUNTRY.SYS

DEVICEHIGH=C:\MOUSE\MOUSE.SYS /1

SHELL=C:\COMMAND.COM /P /E:4096 STACKS=9,256

Содержимое файла autoexec.bat (в машине используется пакет STACKER)

@REM THE CHECK LINE BELOW PROVIDES ADDITIONAL SAFETY

@REM PLEASE DO NOT REMOVE IT.

@C:\STACKER\CHECK /WP

SET COMSPEC=C:\COMMAND.COM

LH C:\DOS\SMARTDRV.EXE 512 256 /V

PATH С:\SIDIAK;C:\;C:\DOS;..C:\STACKER; E:\LEXICON

SET BLASTER= A220 I10 D1 T4

SET CLIPPER=F100

REM БЕЗ NOVCPI - SET CLIPPER=E0

SET TEMP=C:\WINDOWS\TEMP

SET LEX=D:\TEXTKAF

LH C:\UTIL\RUS\UNISCR

LH C:\UTIL\RUS\UNIKBD LH C:\VC\VC

Приложение 3. Пример файла itog.txt с результатами расчетов

|  |  |
| --- | --- |
| Исходная совокупность объектов разбита | на 1 группу. |
| Всего обсчитано | 500 обьектов |
| Выходной признак | ВОЗРАСТ ДЕПУТАТОВ ГОСУДАРСТВЕННОЙ ДУМЫ |
| Параметр классификации | 0.50 |
| Число входных признаков | 3 |
| Входной признак | ОСНОВНАЯ ПРОФЕССИЯ |
| Входной признак | НАЦИОНАЛЬНОСТЬ |
| Входной признак | ПАРТИЙНОСТЬ |

|  |  |  |  |
| --- | --- | --- | --- |
| N групп | Число объектов | Среднее значение выходн. признака | Ст. откл. вых.призн. |
| N 1 | 500 | 42.56 | 9.84 |

ГРУППА N 1

В данную группу входит 500 объектов со следующими номерами:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 102 | 254 | 154 | 157 | 166 | 177 | 321 | 150 | 158 | 160 |
| 161 | 162 | 167 | 169 | 174 | 175 | 195 | 201 | 239 | 240 |
| 241 | 242 | 245 | 250 | 256 | 260 | 263 | 264 | 265 | 267 |

и так далее ...

Выходной признак - ВОЗРАСТ ДЕПУТАТОВ ГОСУДАРСТВЕННОЙ ДУМЫ

Среднее значение выходного признака - 42.56

Стандартное отклонение - 9.84

Группа определяется следующими значениями входных признаков:

Признак N 1 - ОСНОВНАЯ ПРОФЕССИЯ и так далее....

**Список литературы**

1. Устинов В.А., Фелингер А.Ф. Историко- социальные исследования, ЭВМ и математика.-М., 1973.

2. Первоначальная версия программы была реализована на FORTRAN 77 А. Кардаш и А. Бегуном под руководством В. Саакяна.

3. Самохин А.Б., Самохина А.С. Фортран и вычислительные методы для пользователя IBM PC. М., Русина, 1994.