Балашов О.В., Лосева В.А. Извлечение знаний для систем поддержки принятия решений // Международный журнал информационных технологий и энергоэффективности. – 2019. – Т.4 №1(11) с. 10-15

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 681.3.06

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ДЛЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Балашов О.В., Лосева В.А.

Смоленский филиал АО «Радиозавод», Россия, (214027, г. Смоленск, улица Котовского,2), e-mail: smradio@mail.ru

Рассматриваются подход к автоматической кластеризации /классификации объектов по данным обучающей выборки с применением современных инструментальных средств. Результат может быть полезен при проектировании систем поддержки принятия решений.

Ключевые слова: решение, кластеризация, выбор, классификация, обучающая выборка, лингвистическое описание.

EXTRACTION OF KNOWLEDGE FOR DECISION SUPPORT SYSTEMS

Balashov O.V., Loseva V.A.

Smolensk branch of joint-stock company "Radio factory", Russia, (214027, Smolensk, street Kotovskogo, 2), e-mail: smradio@mail.ru

The approach to automatic clusterization (classification) of objects according to learning sampling with application of modern tools is considered. The result can be useful at decision support systems.

Key words: decision, clusterization, choice, classification, training sample, linguistic description.

Качество функционирования системы поддержки принятия решений (СППР) существенно зависит от содержимого её базы знаний. Как известно, существуют две основные группы методов получения знаний: прямые (интервью, изучение литературы и др.) и косвенные (анализ обучающего множества примеров, наблюдения за экспертом и др.) [1, 2]. Проведённые исследования показали, что при принятии решений в условиях неопределённости большую предпочтительность имеют методы второй группы.

В данной статье рассматривается задача автоматической кластеризации по примерам обучающей выборки с выдачей результата в виде совокупности продукционных правил вида «если – то». Решение задачи проводится с использованием инструментальных средств SPSS 13.0 [3] и See5/C5.0 [4].

Постановка задачи.

Имеются массив экспериментальных данных, представленный «примерами» в виде векторов $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = 1,2,\ldots N$, где x_{ij} – некоторые числа ($j = 1,2,\ldots n$), отражающие значения количественных признаков $x_1 \div x_n$, N – общий объем выборки (количество обучающих примеров).

Предполагается, что представленные примеры отражают некоторое, априори неизвестное число m типов различимых объектов, например, различных способов выполнения работ, типов летательных аппаратов, сортов фруктов и т.п.

Требуется: по данным обучающей выборки провести автоматическую кластеризацию представленных примеров по типам объектов, определить число таких типов (кластеров), выделить наиболее информативное подмножество признаков кластеризации, сформулировать решение в виде совокупности отмеченных выше продукционных правил, т.е. в лингвистической форме – для облегчения дальнейшей ручной или полуавтоматической классификации объектов в системах СППР.

Известные методы решения задачи кластеризации. Существует большое количество различных методов решения задачи кластеризации (см., в частности, книги [1, 2, 4-9]), однако в большинстве из них количество кластеров априори задается пользователем, исходя из каких-либо содержательных представлений о характере будущего решения. Практически неизвестны методы, в которых бы, наряду с решением задачи кластеризации проводилась оценка значимости признаков. «Ручные» вычисления по данным методам пригодны лишь для задач небольшой размерности – с числом примеров не более 20, при 2÷3 признаках классификации.

В исследуемой задаче, как число примеров, так и число признаков достаточно велико, что требует привлечения того или иного инструментального (программного) средства, реализующего те или иные алгоритмы кластеризации.

Выбор инструментальных средств. Поскольку существуют инструментальные средства (программы, программные системы), позволяющие решать подобные задачи с помощью персональных компьютеров, метод решения поставленной задачи целиком зависит от выбранного инструментального средства и его возможностей, при этом пользователь математическими деталями используемых алгоритмов может и не интересоваться (эти алгоритмы, как отмечалось, достаточно подробно описаны, например, в монографиях [5, 6]).

В качестве инструментальных средств для решения поставленной задачи в данном случае выбраны:

1) пакет для статистических вычислений SPSS, 13-я версия;

2) программа See5 (версия 1.20a).

Такой выбор поясняется не только широкими возможностями указанных программ, но и тем, что они и правила их использования достаточно подробно описаны в отечественной литературе (см. [3, 4]).

Решение задачи. Предлагаемый подход продемонстрируем на следующем иллюстративном примере.

Пусть имеются объекты двух типов (еще раз оговариваем, что это число предполагается неизвестным), каждый из которых характеризуется двумя числовыми признаками, а соответствующие объектам примеры отображены в таблице 1. Данные подвергнуты рандомизации, т.е. примеры перемешаны случайным образом; в условиях примера – для контроля – принадлежности объекта к тому или иному классу приведены в крайне правом столбце матрицы (они были известны экспериментатору, но неизвестны программе).

Этап 1. Подготовка исходных данных.

Приведенные в таблице исходные данные (10 примеров – т.е. 10 пар значений признаков x_{i1}, x_{i2}) были загружены в таблицу программы SPSS для проведения кластеризации и выявления наиболее информативных признаков.

Этап 2. Выявление числа кластеров и наиболее информативных признаков. После загрузки данных в среду программы SPSS 13.0 дальнейшие исследования базировались на возможности этой программы решать задачу кластеризации несколькими методами, из которых наибольший интерес представляют так называемый метод двухступенчатой кластеризации (TwoStep Cluster). Данный метод, реализованный в системе SPSS 13.0,

позволяет не только автоматически определять оптимальное число кластеров в наборе данных, но и выделять наиболее информативные (с точки зрения задачи кластеризации) признаки.

Таблица 1 – Примеры обучающей выборки

№№ примеров	Признаки		№ класса
	x_1	x_2	
1	9.872	12.406	1
2	11.089	10.268	1
3	-10.19	21.911	2
4	-11.663	21.068	2
5	9.886	10.167	1
6	9.102	9.207	1
7	11.367	10.591	1
8	-8.506	21.161	2
9	-10.006	21.394	2
10	-10.166	20.29	2

С использованием этого метода для исследуемой выборки данных были получены следующие результаты, отраженные в файле отчета программы, фрагменты которого приведены ниже (таблицы 2 и 3).

Двухэтапный кластерный анализ

Таблица 2 – Распределение по кластерам

Распределение по кластерам		N	% объединенных	% от итога
Кластер	1	5	50,0%	50,0%
	2	5	50,0%	50,0%
	Объединенный	10	100,0%	100,0%
Итог		10		100,0%

Таблица 3 – Профили кластеров

Профили кластеров		**Центроиды**			
		x1		x2	
		Среднее	Стд. отклонение	Среднее	Стд. отклонение
Кластер	1	-10,1062	1,11858	21,1648	,58822
	2	10,2632	,94128	10,5278	1,16981
	Объединенный	,0785	10,77977	15,8463	5,67374

Заметим, что программа выдает также информацию об отнесении каждого из примеров обучающей выборки к тому или иному кластеру (классу). Эта информация будет использована при применении второй из рассматриваемых программ.

Как видно, программа правильно выделила два класса (кластера), более того, из её выходных данных следует, что все примеры были классифицированы правильно, а оба признака оказались значимыми (с вероятностью 0,95).

Вторая из таблиц отчёта содержит статистическую информацию о центрах кластеров.

Этап 3. Лингвистическое описание классов. Исследование на данном этапе проводилось с помощью программы See5 [4], которая позволяет по данным экспериментальной выборки (а

также по выявленным для каждого примера номера класса) формировать продукционные правила для лингвистической классификации объектов. Предварительно были подготовлены 2 текстовых файла – с имеющимися данными и именами переменных (файлы **Кластер.names** и **Кластер.data**).

Файл **Кластер.names**

```
class.
class: 1,2.
x1: continuous.
x2: continuous.
```

Файл **Кластер.data**

```
1,9.872,12.406
1,11.089,10.268
2,-10.19,21.911
2,-11.663,21.068
1,9.886,10.167
1,9.102,9.207
1,11.367,10.591
2,-8.506,21.161
2,-10.006,21.394
2,-10.166,20.29
```

В файле **Кластер.data** первые элементы каждой строки отражают принадлежность объекта (примера обучающей выборки) к тому или иному классу, определенному программой SPSS.

Результаты использования программы See5 (отражаемые протоколом в файле **Кластер.out**) приведены ниже.

```
See5 [Release 1.20a]    Tue Sep 12 17:51:27 2006

   Options:
      Rule-based classifiers

Class specified by attribute `class'

Read 10 cases (3 attributes) from Кластер.data

Rules:

Rule 1: (5, lift 1.7)
      x1 > -8.506
      ->  class 1  [0.857]

Rule 2: (5, lift 1.7)
      x1 <= -8.506
      ->  class 2  [0.857]

Default class: 1
```

```
Evaluation on training data (10 cases):

    Rules
----------------
 No     Errors

  2    0( 0.0%)   <<


 (a)   (b)    <-classified as
----  ----
   5          (a): class 1
         5    (b): class 2
```

Интерпретация приведенных результатов такова: всего исследовано 10 случаев, при этом выявлено 2 продукционных правила типа «если-то». Ошибки в классификации отсутствуют. Объединяя правила, можно дать их лингвистическую интерпретацию в виде одного правила:

П: если $x_1 \leq -8.506$, то объект относится к классу 2, иначе – к классу 1.

Отметим, что программа «определила» степень уверенности в справедливости классификации по приведенным правилам 0,857. Небезынтересно заметить, что в данном случае информационно значимым для классификации оказался только один показатель – x_1.

Нетрудно проверить (см. таблицу 1), что в условиях приведенного примера задача выявления продукционных правил решена безошибочно.

Таким образом, автоматически сформулированы продукционные правила, позволяющие по натуральным значениям информативных признаков относить предъявляемый объект к тому или иному классу.

Точность полученного решения следует оценить на уровне 80÷90%, что для многих практических задач следует считать приемлемым.

Следует указать, что к получаемым с помощью предложенного подхода результатам следует относиться с известной долей осторожности (как, впрочем, ко всем статистическим выводам, сделанным на основе только экспериментальных данных), проверяя их, по возможности, другими подходами.

Список литературы

1. Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. – Новосибирск, Изд-во Ин-та математики, 1999. – 212 с.
2. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск, Изд-во Ин-та математики, 1999. – 270 с.
3. Бююль А., Цёфель П. SPSS: искусство обработки информации, анализ статистических данных и восстановление скрытых закономерностей. – СПб.: ООО "ДиаСофтЮП", 2002. – 608 с.
4. Дюк В., Самойленко А. Data mining: учебный курс. – СПб.: Питер, 2001. – 368 с.
5. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.

References

1. Lbov G. S., Startseva N.G. Logic decision functions and questions of statistical stability of decisions. - Novosibirsk, Publishing house In mathematicians, 1999. (in Russian)
2. Zagorujko N.G. Applied methods of the analysis of data and knowledge. Novosibirsk, Publishing house In mathematicians, 1999. (in Russian)
3. Buul A., Cefel P. SPSS: art of processing of the information, the analysis of statistical data and restoration of the latent laws. - SPb.: Open Company "DiaSoftUP", 2002. (in Russian).
4. Duk V., Samoilenko A. Data mining: a training course. - SPb.: Peter, 2001. (in Russian).
5. Osovsky S. Nejronnye of a network for information processing. - M: the Finance and statistics, 2002. (in Russian).