基于决策树的模糊聚类评价算法及其应用

王园园[1,2],倪志伟[1,2],赵裕啸[1,2],伍章俊[1,2]
(1.合肥工业大学 管理学院,安徽 合肥 230009;
2.合肥工业大学 过程优化与智能决策教育部重点实验室,安徽 合肥 230009)

摘 要:评判聚类结果的有效性是一个复杂问题。文中提出一种基于决策树的模糊聚类评价算法,证明了决策树过程实际上就是一种模糊聚类评价过程,因此能够使用决策树算法来评价聚类结果的好坏,并在此基础上提出一个新的定理。通过UCI中已经有准确聚类个数的数据来验证算法有效性,在实验中首先使用K-means聚类算法得到不同聚类结果,再使用决策树中的C5.0算法来评价各种聚类结果,得到的最优聚类结果与UCI数据原有的聚类结果接近,证明了算法的实用性。最后给出算法在证券行业客户细分中的应用实例。

关键词:聚类;模糊评价;决策树;客户细分
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)09-0232-04

Fuzzy Clustering Evaluation Algorithm Based on Decision Tree and Application

WANG Yuan-yuan[1,2], NI Zhi-wei[1,2], ZHAO Yu-xiao[1,2], WU Zhang-jun[1,2]
(1. School of Management, Hefei University of Technology, Hefei 230009, China;
2. Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision-Making, Hefei 230009, China)

Abstract: The evaluation of clustering result is a comprehensive problem. Proposed a new clustering evaluation algorithm based on decision tree, it verified the decision tree algorithm actually was a fuzzy clustering evaluation process, so it can use decision tree algorithm to evaluate clustering result, then proposed a new theorem on the basis. It used UCI data with precise clustering result to verify the efficiency of the combination algorithm, in the experiment it first used K-means algorithm to get different clustering results, then used C5.0 algorithm to evaluate these results, the best result is close to original clustering result, so it demonstrates the efficiency of the algorithm. At last it gave a practical example of customer segmentation in security.

Key words: clustering; fuzzy evaluation; decision tree; customer segmentation

0 引 言

聚类过程可以看作是一种无监督的学习过程,因为没有预先定义的分类或示例来表明数据集中哪种期望的关系是有效的,如何用一种客观公正的质量评价方法来评判聚类结果的有效性是一个困难而复杂的问题。文献[1]中介绍了多种聚类算法的评价方法,如Cophenetic相关系数法是基于比较矩阵的相似程度,而SD有效性指数则是基于聚类平均散布性和聚类间总体分离性的一种相对度量方法,这些方法通常是将点划分给固定的类,但是对于实际中数据,其分类通常带有模糊性。

模糊聚类有效性问题的研究主要集中在模糊C-均值(FCM)算法的两类有效性函数上:

①有效性函数将隶属度作为有效性评价的主要因素,像Bezdek提出的划分系数PC[2]和划分熵PE[3]都属于这一类;

②有效性函数不仅仅将隶属度考虑在内,还考虑到了数据集本身,Xie-Ben指标[3]就是这一类有效性函数的典型代表。

文中提出一种基于聚类与决策树结合的细分模型,利用决策树与聚类有效性模糊评价法的相似性,通过决策树完成对聚类结果进行微观评测和提炼分类规

收稿日期:2008-12-27;**修回日期**:2009-02-25
基金项目:国家高技术研究发展计划(863)(2007AA04Z116);国家自然科学基金(70871033)
作者简介:王园园(1984-),女,硕士研究生,研究方向为数据挖掘与数据流;倪志伟,教授,博士生导师,研究方向为人工智能,机器学习。

则,使用 UCI[4]数据评价分类结果,最后完成客户细分。

1 相关概念

1.1 聚类算法

聚类中著名的有 K-Means[5]、K-Center[6]、DBSCAN[7]和 OPTICS[8]等。其中 K-Means 聚类算法由 Mac Queen[9]提出,具有算法结构简单、收敛速度快的优点,适用于大规模的数据分析。但 K-Means 具有两大缺点:一是必须事先已知或者给定簇的个数;二是聚类结果受初始聚类中心影响较大。

1.2 决策树算法

决策树[10]常见算法有 ID3[11],C4.5[12],C5.0[12]等。在 C5.0 算法中使用信息论方法对大量实例的特征进行信息量分析,基于信息熵的方法递归形成决策树,以计算属性 X 为例计算它的信息增益率 GainRatio(X)。S 表示一组样本,p_i 是任意样本属于 D_i 的概率,用 S_i/S 表示。假定类别属性具有 n 个不同的值,定义 n 个不同类 $D_i(i=1,\cdots,n)$。设 S_i 是类 D 中的样本数。Info(S) 表示当前样本中的信息熵,计算如下:

$$\text{Info}(S)=-\sum_{i=1}^{n}p_i\log(p_i) \tag{1}$$

设属性 X 具有 n 个不同值$\{X_1,X_2,\cdots,X_n\}$,利用 X 将 S 划分为 n 个子集$\{S_1,S_2,\cdots,S_n\}$,其中 S_j 为 S 中在 X 中具有 X_j 的样本,S_{ij} 是子集 S_j 中类 D_i 样本数。Info(S,X) 表示利用属性 X 划分 S 中所需要信息熵,计算如下:

$$\text{Info}(S,X)=\sum_{j=1}^{n}\frac{S_j}{S}\text{Info}(S_j) \tag{2}$$

分裂信息 SplitInfo(X) 是 S 关于属性 X 的各值的熵,用以消除具有大量属性值属性的偏差,计算如下:

$$\text{SplitInfo}(X)=-\sum_{i=1}^{n}\frac{|S_i|}{|S|}\log\left(\frac{|S_i|}{|S|}\right) \tag{3}$$

$$\text{Gain}(X)=\text{Info}(S)-\text{Info}(S,X) \tag{4}$$

$$\text{Gain}(X)=(\text{Info}(S)-\text{Info}(S,X))/\text{SplitInfo}(X)$$

1.3 模糊聚类评价

模糊聚类评价[1]是寻求数据集中大多数矢量在一个簇中有高的隶属度的聚类模式。一个模糊聚类由隶属度矩阵 $\boldsymbol{P}=[p_{ij}]$表示。这里 p_{ij}代表簇 j 中矢量 i 的隶属程度。则关键指标划分系数

$$PC=N\sum_{i=1}^{N}\sum_{j=1}^{c}p_{ij}^2$$

其中:c 为聚类个数;PC 取值为$[1/c,1]$。PC 越接近 1,划分越清晰;反之,PC 越接近 $1/c$,划分越模糊。

关键指标划分熵为

$$PE=-N\sum_{i=1}^{N}\sum_{j=1}^{c}p_{ij}\times\log(p_{ij}) \tag{5}$$

PE 取值为$[0,\log c]$,$c>1$。分类越分明时,PE 的值就越小;分类越模糊时,PE 的值就越接近于 $\log c$。

2 基于决策树的模糊聚类评价算法

2.1 决策树算法与模糊聚类评价算法联系

模糊聚类评价方法与决策树算法均是采用信息熵来评价隶属度,而信息熵之间存在相关性,可以得出模糊聚类评价方法与决策树算法也存在相关性,因此在文中提出一个新的定理:

定理 1　聚类模糊评价法划分熵的大小随决策树信息熵增益的增大而减少。

证明:由式(5)条件得 $c>1$,不妨设其为 N 即分类的个数,则式(5) 转换为向量积:

$$\begin{aligned}PE&=-N\sum_{i=1}^{N}\sum_{j=1}^{N}p_{ij}\times\log(p_{ij})\\&=-N\{\sum_{j=1}^{N}p_{1j}\times\log(p_{1j})+\sum_{j=1}^{N}p_{2j}\times\log(p_{2j})+\cdots\\&\quad+\sum_{j=1}^{N}p_{nj}\times\log(p_{nj})\}\\&=\left(\sum_{j=1}^{N}p_{1j}\times\log(p_{1j}),\sum_{j=1}^{N}p_{2j}\times\log(p_{2j}),\cdots,\sum_{j=1}^{N}p_{nj}\times\log(p_{nj})\right)\cdot(1,1,\cdots,1)\cdot(-N)\end{aligned} \tag{6}$$

由式(1) 和式(2) 得:

$$\text{Info}(S,X)=\sum_{j=1}^{n}\frac{S_j}{S}\text{Info}(S_j)=-\sum_{j=1}^{n}\frac{S_j}{S}\sum_{i=1}^{n}p_{ij}\times\log(p_{ij})$$

该公式的各符号含义为:分类目标属性有 N 种不同值$(1,2,3,\cdots,i,\cdots,n)$,原数据集按属性 X 可划分为 N 类$\{S_1,S_2,\cdots,S_j,\cdots,S_n\}$,其中 S_j 类中分类目标属性值为 i 的概率为 p_{ij}。

因为 i、j 取值范围相同,则可以互换 i、j 得:

$$\begin{aligned}\text{Info}(S,X)&=-\sum_{j=1}^{n}\frac{S_j}{S}\sum_{i=1}^{n}p_{ij}\log(p_{ij})\\&=-\sum_{i=1}^{n}\frac{S_i}{S}\sum_{j=1}^{n}p_{ij}\log(p_{ij})\\&=-\left(\frac{S_1}{S}\sum_{j=1}^{N}p_{1j}\times\log(p_{1j})+\frac{S_2}{S}\sum_{j=1}^{N}p_{2j}\times\log(p_{2j})\right.\\&\quad\left.+\cdots+\frac{S_n}{S}\sum_{j=1}^{N}p_{nj}\times\log(p_{nj})\right)\\&=-\left(\sum_{j=1}^{N}p_{1j}\times\log(p_{1j}),\sum_{j=1}^{N}p_{2j}\times\log(p_{2j}),\cdots,\sum_{j=1}^{N}p_{nj}\times\log(p_{nj})\right)\cdot\partial\end{aligned}$$

其中系数向量 $\partial = (\frac{S_1}{S}, \frac{S_2}{S}, \cdots, \frac{S_n}{S})$

因为$(1,1,\cdots,1) \cdot (1,1,\cdots,1) \times \frac{1}{N} = 1$

$$\mathrm{Info}(S,X) = -\left(\sum_{j=1}^{N} p_{1j} \times \log(p_{1j}), \sum_{j=1}^{N} p_{2j} \times \log(p_{2j}), \cdots, \sum_{j=1}^{N} p_{nj} \times \log(p_{nj})\right) \cdot \partial \cdot (1,1,\cdots,1) \cdot (1,1,\cdots,1) \times \frac{1}{N}$$

$$= -N\left(\sum_{j=1}^{N} p_{1j} \times \log(p_{1j}), \sum_{j=1}^{N} p_{2j} \times \log(p_{2j}), \cdots, \sum_{j=1}^{N} p_{nj} \times \log(p_{nj})\right) \cdot (1,1,\cdots,1) \cdot (1,1,\cdots,1) \cdot \partial \times \frac{1}{N^2}$$

$$= PE \cdot (1,1,\cdots,1) \cdot \partial \times \frac{1}{N^2}$$

可得 $\mathrm{Info}(S,X) = PE \cdot \alpha$

其中 $\alpha = (1,1,\cdots,1) \cdot \partial \times \frac{1}{N^2}$ 为一正实数。

设 $\beta = 1/\alpha$

$$PE = \mathrm{Info}(S,X) \cdot \beta \tag{7}$$

即 PE 与 $\mathrm{Info}(S,X)$ 存在正比例关系

由式(4) 得

$$\mathrm{Info}(S,X) = \mathrm{Info}(S) - \mathrm{Gain}(X) \tag{8}$$

则由(6)、(7)、(8) 三式可得:

$$PE = [\mathrm{Info}(S) - \mathrm{Gain}(X)] \cdot \beta$$

易知在数据集确定情况下,β 为定值且都为正值,$\mathrm{Info}(S)$ 与 $\mathrm{Gain}(X)$ 也为正值,则当信息增益 $\mathrm{Gain}(X)$ 增大时 PE 减小。证毕。

由定理可知聚类模糊评价法划分熵的大小随决策树信息熵增益的增大而减少,划分熵减少则表示聚类更有效,而信息熵增益越大表示决策树分类越好,该分类属性也将越重要且更可能被首先使用。那么可以得出在决策树执行过程中,实际上已经对聚类结果的进行了检测,因此决策树过程实际就是一种模糊聚类评价过程。

通过采用 UCI 提供的机器学习数据库中 Hand poker DataSet 进行测试,检验方法的可行性。数据集原分类个数为 4,使用决策树聚类组合算法评测后,4 也为最佳聚类个数,正确率见表 1。

表 1 UCI 数据准确表

类别	正确聚类所含数据个数	算法实验所含数据个数	属于正确分类的个数	分类准确率	分类误判率
1	1789	1690	1600	89.4%	5.0%
2	128	139	128	100%	8.5%
3	756	820	650	85.9%	8.5%
4	357	381	310	84.0%	19.8%

2.2 基于决策树的模糊聚类评价算法步骤

基于决策树的模糊聚类评价算法步骤如下:

(1)应用 K-Means 将数据集分别进行 8 种不同的划分,每种簇的个数分别为 A~H。

(2)抽取数据与决策树过程。以划分 A 个簇为例,对 A 个簇中数据随机抽取,获得训练集 A1 及测试集 A2,C5.0 算法则完成对测试集中客户类别的预测。

(3)聚类评价。通过观察决策树规则所覆盖的数据占 A1 的比率,以及应用规则对测试集 A2 中数据进行预测准确性的高低,得到最好的聚类方法。

3 基于决策树模糊聚类评价算法应用过程

3.1 证券业客户数据采集与处理

证券公司的数据库中存有所有客户的交易及个人资料。实验中主要收集股民的行为属性,包括年初资产,年末资产,资产盈亏率,交易次数,存取次数,存取差额。

3.2 实验步骤

实验步骤如下:

(1)K-Means 聚类方法。

将最后处理得到样本集数据客户最终数据表分成 2~9 个类,首先进行 2 个类的初始划分。聚类完成之后,按照客户资金帐号给每个客户标上标签,即 1,2 两种标签。客户最终数据表中共有 188862 条客户数据,选择 165277 作为接下来决策树的训练集,23585 作为测试集。依据相同的原理,得到 3~9 个分类中各自不同的训练集及测试集。

(2)C5.0 决策树算法的运用。

得到 8 组不同的训练集及测试集之后,使用 C5.0 算法生成 8 个决策树。

(3)两种比较之后,选择出两个值都相对较好的聚类方法,结果如图 1、2 所示。

在对训练集进行规则一及规则二的检测后,得到分六类效果最好。

3.3 实验结果分析

实验结果见表 2。

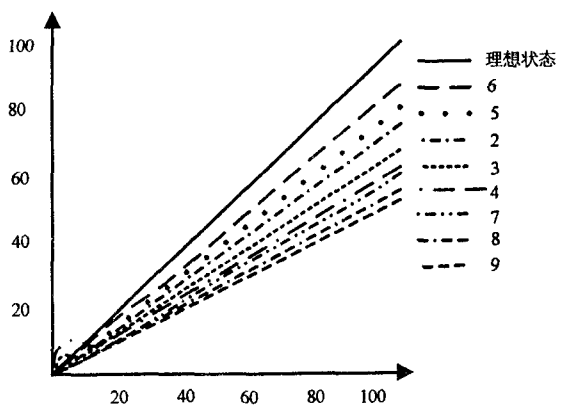

图 1 证券数据训练集准确图

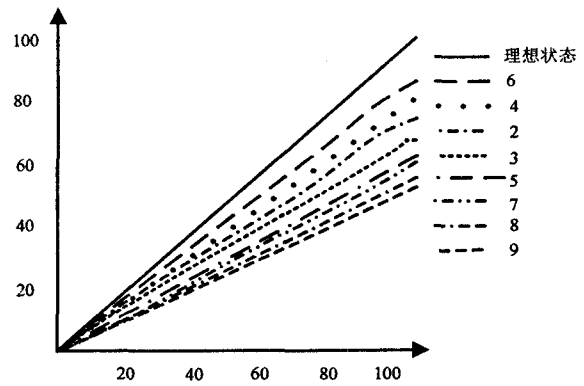

图 2　证券数据预测集准确图

表 2　客户分类信息表

类别	人数	佣金贡献(平均)	年度盈亏率(平均)	交易次数(平均)
1	62095	145.78	+1.01	23.67
2	57295	263.12	-1.32	42.63
3	25691	1047.37	+2.53	93.29
4	35402	5943.23	-4.56	473.83
5	7594	23.69	+1.47	10.56
6	785	22127.85	+5.39	788.63

类别	存取次数(平均)	存取差额(平均)	年初资产(平均)	年末资产(平均)
1	10.06	3538.69	28787.29	32356.78
2	24.30	27896.65	76875.63	96503.37
3	3.37	2564.33	35896.79	107236.45
4	73.42	154654.56	65643.78	75677.33
5	33.21	569.56	1230.54	2089.21
6	129.21	585466.63	2678344.95	10237936.36

其中存取差额、年初资产、年末资产计量单位为元。

分析各类客户特征为:

(1)资金量较少,交易及存取次数少,基本做到不赔不赚,年末、年初资产都较小,客户个体佣金贡献较少,但是整体佣金贡献很大。属于中小散户。

(2)买卖及资金存取有提升,年初、年末资产较高,佣金贡献超过第一类。所占人数多,因为亏损,有可能流失。

(3)交易次数明显提高,资金存取量较少,年初资产较少,但是年末资产较高,赢利率较高,佣金贡献也明显增多。对市场敏感,专业知识基础好。

(4)亏损率很高,资金存取量很多,在股市中追加了大量资金,资金实力较高,交易次数很频繁。缺乏专业知识,迫切解套,佣金贡献虽然高,但很有可能流失。

(5)各项指标都较低,有少量的赢利,属于试探性客户。

(6)年初,年末资产都很大,交易次数很频繁,盈利率很高,佣金贡献很大,属于质量很高的客户。经济实力很高,对市场很敏感,操纵能力强。

4　结束语

提出了基于决策树的模糊聚类评价算法模型,通过 C5.0 决策树算法完成模糊聚类评价。通过研究,可以发现聚类决策树组合算法在客户细分中的重要应用。

文中提出的算法具有一定的实用性,但还需要在实际客户细分系统中作进一步检验,并使之不断完善。

参考文献:

[1] Halkidim, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Intelligent Information Systems, 2001, 17(2-3):107-145.

[2] Bezdek J C. Numerical Taxonomy with Fuzzy Sets[J]. J Math Biol, 1974, 1(1):57-71.

[3] Bezdek J C. Cluster Validity with Fuzzy Sets[J]. Journal of Cybernetics, 1974, 3(3):58-73.

[4] Hettich S, Blake C L, Merz C J. UCI Repository of Machine-Learning Databases[EB/OL]. 1998-01. http://www.ics.1uci.edu/mlearn/MLRepository.html.

[5] 梁　循. 数据挖掘:建模、算法、应用和系统[J]. 计算机技术与发展, 2006, 16(1):1-4.

[6] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000:200-245.

[7] Ester M, Kriegel H P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases [C]//Proc. of 1996 Intl. Conf. on Knowledge Discovery and Data Mining. Portland OR:[s.n.], 1996:226-231.

[8] Ankerst M, Breuning M, Kriegel H P, et al. Optics: Ordering Points to Identify the Clustering Structure[C]//Proc. of 1999 ACM - SIGMOD Intl. Conf. on Management of Data. Philadelphia, PA:[s.n.], 1999.

[9] Mac Queen J. Some methods for classification and analysis of multivariate observations[C]//Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley:[s.n.], 1967.

[10] 谢寰红. 数据挖掘在证券公司 CRM 客户细分中的应用[J]. 计算机工程, 2004(1):553-554.

[11] Ruggieri S. Efficient C4.5[J]. IEEE Transactions on Knowledge and Data Engineering, 2002(4):438-445.

[12] 赵建华, 陈汉林, 杨树锋, 等. 基于决策树算法的滑坡危险性区划评价[J]. 浙江大学学报:理学版, 2004, 31(4):465-470.