

文章编号:1001-9081(2010)07-1926-04

修正核函数模糊聚类算法

赵国亮, 黄沙日娜

(黑龙江科技学院 数力系, 哈尔滨 150027)

(ocnzhao@gmail.com)

摘要:应用核函数度量的紧致性和分离性,给出了一种新的聚类有效性指标 KKW,由 KKW 指标得到最优聚类数并用于修正核函数模糊聚类算法(MKFCM),由于经过了修正核函数的映射,使原来没有显现的特征突显出来。用 MKFCM 对 Wine 和 glass 数据集进行聚类,每一类的聚类正确度大于 90%;对于缺失数据的 Wisconsin Breast Cancer 数据,错分率为 4.72%。该聚类方法在性能上比经典聚类算法有所改进,具有更快的收敛速度以及较高的准确度。仿真实验的结果证实了修正核聚类方法的可行性和有效性。

关键词:模糊 C 均值算法; 模糊聚类; 核函数; 有效性指标; 聚类个数估计

中图分类号: TP18 文献标志码:A

Fuzzy clustering algorithm with modified kernel functions

ZHAO Guo-liang, HUANG Sha-rina

(Department of Mathematics and Mechanics, Heilongjiang Institute of Science and Technology, Harbin Heilongjiang 150027, China)

Abstract: Using kernelized metric of compactness and separation, this paper proposed a new clustering validity index named KKW, and obtained the optimized cluster number. Besides, the KKW index was used in the modified kernel fuzzy clustering (MKFCM) algorithm. As mapped by modified Mercer kernel functions, the data set shows new features never showed before. MKFCM algorithm was applied to the data set Wine and glass. For every clustered class, MKFCM has overall accuracy higher than 90%; as to the incomplete data set Wisconsin Breast Cancer, difference is 4.72%. The modified kernel clustering algorithm is faster than the classical algorithm in convergence and more accurate in clustering. The results of simulation experiments show the feasibility and effectiveness of the modified kernel clustering algorithm.

Key words: Fuzzy C-Mean (FCM) algorithm; fuzzy clustering; kernel function; validity index; clusters number estimation

0 引言

基于模糊集理论的模糊聚类方法是根据样本点与聚类中心的隶属度大小来划分归属类别的,其中以模糊 C 均值(Fuzzy C-Mean, FCM)算法应用最为广泛,它被广泛应用于很多领域,如:模式识别、图像处理、基因分类以及计算机视觉等。最近,许多文章研究了应用核函数^[1-5]进行聚类的方法,这些方法首先把数据映射到一个高维空间中,以便能获得更好的线性可分性,然后计算样本在输入空间的相似性度量。实验证明,核模糊 C 均值(Kernelized Fuzzy C-Mean, KFCM)优于传统的 FCM 算法^[3-8]。本文提出一种新的聚类有效性指标 KKW,并利用此指标计算出最优聚类数目,然后利用新提出的修正核距离函数模糊聚类算法给出聚类结果。

1 修正核函数

核是一个函数 $K(\mathbf{x}, \mathbf{y})$, 描述了在某个特征空间 \mathbf{H} 中的一个内积,对所有 X 中的 \mathbf{x}, \mathbf{y} , 满足 $K(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x})\varphi(\mathbf{y})$ 。其中 φ 是从 X 到特征空间 \mathbf{H} 的映射。

定理 1^[3] $K(\mathbf{x}, \mathbf{y})$ 表示一个连续的对称核, $\mathbf{x}, \mathbf{y} \in X$, 核 $K(\mathbf{x}, \mathbf{y})$ 可以展开为:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x})\varphi_i(\mathbf{y}); \lambda_i > 0 \quad (1)$$

式(1)绝对一致收敛的充要条件是 $\int_a^b \int_a^b K(\mathbf{x},$

$y)g(\mathbf{x})g(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0$ 对于所有满足条件的 $g(\mathbf{x})(\int g(\mathbf{x})d\mathbf{x} < \infty, g(\mathbf{x}) \neq 0)$ 成立。函数 $\varphi_i(\mathbf{x})$ 称为展开的特征函数, λ_i 称为特征值。

任何一个函数只要满足 Mercer 定理条件就可以作为 Mercer 核。对输入空间的样本集 $\mathbf{x}^{(q)}$, 表示为 $\mathbf{x}^{(q)} \in \mathbf{R}^N (q = 1, 2, \dots, Q)$ 。用一非线性函数 φ 把所有样本映射到高维空间 \mathbf{H} 中,可以得到新的样本集 $\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_N(\mathbf{x})$, 那么,由定理 1 就可以得到:

$$\|\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)\|^2 = (\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j))(\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j) \quad (2)$$

特征空间中的点积就能用输入空间的核来表示,由此可以定义式(3):

$$d(\mathbf{x}, \mathbf{y}) = \|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| \quad (3)$$

$d(\mathbf{x}, \mathbf{y})$ 是特征空间中的欧氏距离,核代入技巧使得在原输入空间中诱导出了一类依赖于核的新的距离度量,由此将 FCM 推广为同一空间中不同距离度量的一种新的聚类方法。引入核函数之后,特征空间的内积运算转化为样本空间中核函数的计算,而不用知道 $\varphi(\mathbf{x})$ 的具体形式。常用的 Mercer 核有:

- 1) 线性核 $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$;
- 2) 多项式核 $k(\mathbf{x}, \mathbf{y}) = (\beta \cdot \mathbf{x} \cdot \mathbf{y} + r)^d, d \in \mathbf{Z}$;
- 3) 高斯核函数 $k(\mathbf{x}, \mathbf{y}) = \exp(-\beta \|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2)$, 其中

收稿日期:2009-12-16;修回日期:2010-02-24。基金项目:黑龙江省教育厅科学技术研究项目(11544048)。

作者简介:赵国亮(1982-)男,内蒙古乌兰察布人,讲师,硕士研究生,主要研究方向:模糊数学、智能优化算法; 黄沙日娜(1982-),女,内蒙古包头人,讲师,硕士研究生,主要研究方向:运筹学、控制论。

σ 为高斯函数的宽度;

4) Sigmoid 函数 $k(\mathbf{x}, \mathbf{y}) = \tanh(-\beta(\mathbf{x} \cdot \mathbf{y}) + r)$, r, β 是自定义参数。

核函数的选择应根据实验来确定。

1995 年, Cortes 和 Vapnik 引入了支持向量机 (Support Vector Machines, SVM)^[11]。在解决某些问题时支持向量机优于其他算法。支持向量机的成功使得核函数被广泛应用于学习算法中^[5]。

定义 1^[1] 对于一个正标量函数 $D(\mathbf{x})$, 定义 $\bar{k}(\mathbf{x}, \mathbf{x}') = D(\mathbf{x})k(\mathbf{x}, \mathbf{x}')D(\mathbf{x}')$, 称之为核函数通过因子 $D(\mathbf{x})$ 的保形变换。 $\bar{k}(\mathbf{x}, \mathbf{x}')$ 称为支持向量机的修正核函数。

定义 2 令 $D(\mathbf{x}) = \frac{1}{1 + k^2 \exp(-\|\mathbf{x} - \mathbf{v}\|^2 / \sigma^2)}$, \mathbf{v} 为聚类中心, 由保形变换 $D(\mathbf{x})$ 所得到的修正核函数 $\bar{k}(\mathbf{x}, \mathbf{x}')$ 满足 Mercer 条件。

证明 设 A 为 \mathbf{R}^n 的紧子集, 对任意 $h(\mathbf{x}) \in L_2(A)$, 有 $\int h^2(\mathbf{x}) d\mathbf{x} < \infty$, 因为 $D(\mathbf{x})$ 是正的标量函数, $k(\mathbf{x}, \mathbf{x}')$ 对称连续, 所以 $\bar{k}(\mathbf{x}, \mathbf{x}')$ 是对称连续的。又 $D(\mathbf{x}) > 0$, 则必存在一个正数 β , 使得 $D(\mathbf{x}) \geq \beta > 0$ 且有:

$$\begin{aligned} & \iint_{A \times A} \bar{k}(\mathbf{x}, \mathbf{x}') h(\mathbf{x}) h(\mathbf{x}') d\mathbf{x} d\mathbf{x}' = \\ & \iint_{A \times A} D(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') D(\mathbf{x}') h(\mathbf{x}) h(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq \\ & \beta^2 \iint_{A \times A} h(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') h(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0 \end{aligned}$$

2 聚类有效性指标

常用的聚类算法是 FCM 算法^[9-10], 该算法要求在聚类前事先指定聚类数 K_0 。通常情况下, 聚类数 K_0 是未知的, 判断聚类数 K_0 的合理性属于聚类有效性问题, 常用来评估聚类有效性的指标有划分系数(Partition Coefficient, PC)^[9] 和划分熵(Partition Entropy, PE)^[10], 这些指标与数据集的几何结构无关。文献[12]提出了 XB 指标:

$$XB = \sum_{i=1}^c \sum_{j=1}^n U_{ij}^m \|x^{(j)} - x^{(i)}\|^2 / nD_{\min} \quad (4)$$

其中 $m = 2$, $D_{\min} = \min_{i \neq j} \|v^{(j)} - v^{(i)}\|^2$ 是类间的最小距离(分离性), U_{ij} 是 $x^{(i)}$ 隶属于 $v^{(j)}$ 的模糊隶属度。Kwon 有效性指标(KW)稍优于 XB 指标。定义为:

$$KW = \sum_{q=1}^c \sum_{k=1}^n U_{qk}^m \|x^{(q)} - c^{(k)}\|^2 + \sum_{k=1}^K \|c^{(k)} - \mu\|^2 / KD_{\min} \quad (5)$$

文献[8]又利用划分系数和指数分离性定义了 PCAES (Partition Coefficient And Exponential Separation) 指标。

$$PCAES = \sum_{j=1}^n U_{ij}^2 / \mu_M - \exp(-D_{\min} / \beta_T) \quad (6)$$

在上述指标中, 紧致性被用来评价类内的内聚程度, 分离性被用来评价类间的分离程度。本文利用核函数提出了一种新的有效性指标 KKW:

$$KKW(c) = (J_s + D_c) / nD_{\min} \quad (7)$$

其中 $J_s = \sum_{i=1}^c \sum_{j=1}^n U_{ij}^m \exp(-\beta \|x^{(j)} - x^{(i)}\|^2 / 2\sigma^2)$, $D_c = \exp(-\beta \sum_{i=1}^c \|x^{(i)} - \mu\|^2 / 2\sigma^2)$, 参数 $\beta = 2$, $\sigma^2 = 10$, $\mu = \frac{1}{c} \sum_{i=1}^c x^{(i)}$ 。有效性指标综合考虑了类内的离差程度和类心间的

的离差程度。 J_s 表示各类点间的核距离之和, D_c 表示类与类心均值的核距离。好的分类应该是类内的点集尽可能紧凑, 类与类间的距离尽可能大。对经典的 glass 和 IRIS 数据集, 不同聚类数下的 KKW 聚类有效性指标如图 1 和图 2。

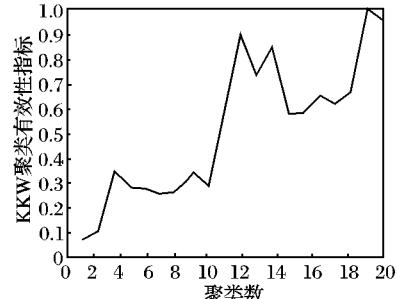


图 1 glass 聚类指标

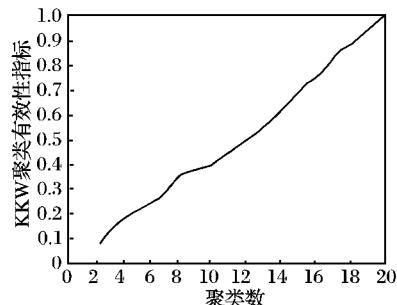


图 2 IRIS 聚类指标(去掉聚类数为 1 的情况)

从图 1 和 2 中可以看到, 用 KKW 划分聚类时, glass 数据集的较小分类有 $c = 2$ 和 $c = 6$, 根据数据集的特征, 划分分别取 $c^* = 6$ 较为合理。对于 IRIS 数据集来说, 有的文献认为划分为 3 类较为合理, 而文献[13-16]均认为是 2 类。由本文提出的 KKW 指标来看, 聚类数取为 $c^* = 2$ 较合理。

3 算法描述

在处理 N 维数据集 $\{\mathbf{x}^{(q)} : q = 1, \dots, Q\}$ 时, 其中 $\mathbf{x}^{(q)} = (x_1^{(q)}, \dots, x_N^{(q)})$, 通常要对数据进行预处理。本文采用文献[5]方法, 对 $\mathbf{x}^{(q)}$ 中每一分量 $x_n^{(q)}$ ($n = 1, 2, \dots, N$) 作线性映射:

$$x_n^{(q)} = (x_n^{(q)} - a_n) / (b_n - a_n); q = 1, \dots, Q \quad (8)$$

其中 $a_n = \min_{q=1, \dots, Q} \{x_n^{(q)}\}$, $b_n = \max_{q=1, \dots, Q} \{x_n^{(q)}\}$ 。通过上述线性映射, 特征属性的取值都在 $[0, 1]$ 内。

3.1 修正核聚类算法(MKFCM)

步骤 1 设定聚类数目 c 、参数 m 和允许误差 E_{\max} 。

步骤 2 初始化聚类中心 \mathbf{v}_i 。

步骤 3 重复下面运算, 直到误差小于允许误差 E_{\max} :

1) 用保形变换 $D(\mathbf{x})$ 修正 $K(\mathbf{x}_k, \mathbf{v}_i)$:

$$K(\mathbf{x}_k, \mathbf{v}_i) = D(\mathbf{x}_k)K(\mathbf{x}_k, \mathbf{v}_i)D(\mathbf{v}_i)$$

$$2) U_{ik} = \frac{(1/(2 - \bar{K}(\mathbf{x}_k, \mathbf{v}_i)))^{1/(m-1)}}{\sum_{i=1}^c (1/(2 - \bar{K}(\mathbf{x}_k, \mathbf{v}_i)))^{1/(m-1)}} \quad (9)$$

3) 更新聚类中心:

$$\mathbf{v}_i = \frac{\sum_{i=1}^n U_{ik}^m \bar{K}(\mathbf{x}_k, \mathbf{v}_i) \mathbf{x}_k}{\sum_{i=1}^n U_{ik}^m \bar{K}(\mathbf{x}_k, \mathbf{v}_i)} \quad (10)$$

3.2 模糊连接矩阵

对数据预聚类后, 得到样本集的一个预聚类结果, 这时的聚类数目 c 远大于实际的最优聚类数目 c^* , 需要进一步把 c 类结果合并为 c^* 类。本文采用如下处理方法:

步骤 1 对每一聚类, 计算其中含有的样本数并按样本

数从大到小排序,从中选出含有样本较多的 c^* 个类,并计算这 c^* 个类的聚类中心(去掉一些野值,由参数 $alpha$ 控制, $alpha = 5$)。

步骤2 利用式

$$y_{ij} = \exp(-\beta \|c^{(j)} - c^{(i)}\|^2 / 2\sigma^2) \quad (11)$$

计算其余 $c - c^*$ 个预聚类的样本到上述 c^* 个聚类中心的平均距离,组成模糊连接矩阵:

$$Y = (y_{ij})_{c^* \times (c-c^*)} \quad (12)$$

步骤3 根据 Y 的隶属度选择相应的分类。

综上所述,根据上面对 MKFCM 算法和模糊连接矩阵的描述,自适应地修正核聚类算法描述如下。

步骤1 用式(1)对原始数据集 data 进行数据预处理,取值于闭区间 $[0,1]$ 。

步骤2 计算最优聚类数目 c^* 。

1)令聚类数 $c = 2, \dots, C_{max}$ 。

2)用 MKFCM 算法计算 U_{ij} 。

3)计算 J_c, D_c 和 D_{min} ,代入式(7)求对应类 c 的聚类有效性指标 $KKW(c)$ 。

4)从聚类有效性指标 $KKW(c)$ 中找出最优聚类数 c^* 。

步骤3 初始化修正核聚类算法参数:样本数 Q ,样本 $\{x^{(q)}, q = 1, 2, \dots, Q\}$,样本特征维数 N ,最优聚类数 c^* ,高斯核函数参数 σ 。

步骤4 对原始数据集进行预聚类。设定预聚类数目 $PreClust$ ($PreClust$ 应远大于最优聚类数 c^*),使用 MKFCM 算法进行聚类。

步骤5 使用式(11)计算模糊连接矩阵 $Y = (y_{ij})_{c^* \times (c-c^*)}$,其中 $\beta = 1$ 。根据 Y 中隶属度最大原则,把这 $c - c^*$ 类合并为 c^* 个类。

4 仿真实验

本文选取了一个人工高斯分布的数据集和两个经典数据集(Wine 和 IRIS)来进行仿真实验。实验在 Matlab 2008b 软件环境下完成,计算机 CPU 为 AMD Athlon4000+,内存 2 GB。

样本集由 5 个高斯分布的类别组成,如图 3 所示。五类的类中心分别是 $(5,5), (5,15), (15,15), (15,5), (10,10)$,它们的方差分别是 $2.5, 2.5, 2.5, 2.5$ 和 1.2 ,五类样本在图 3 中用五种形状的标记区分。样本总数为 250 个,每类 50 个。运行算法 10 次取其平均聚类结果。参数设定为 $m = 2$, $C_{max} = 20$, 预聚类数设定为 $PreClust = 70$ 。运行 MKFCM 进行聚类,得到的聚类图如图 3。

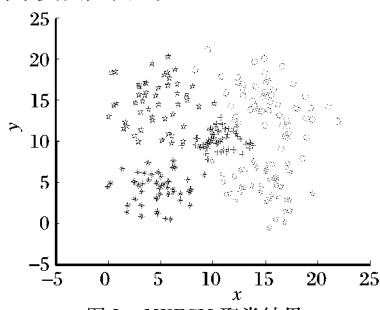


图 3 MKFCM 聚类结果

针对经典数据集(Wine 和 IRIS),算法运行 10 次取平均聚类正确度。参数设定同上。

类 3 的错分样本集较大是因为第三类被其余四类包围且距离较近,容易聚类到其他类别所致。

对于样本集 Wine 来说,MKFCM 的错分率较低。而对于 IRIS 数据集来说,150 个数据,FCM 算法把第一类的 8 个数据错分到了第二类中。KFCM 是 7 个,而 MKFCM 正确地把数据

分为两类:1~50 的样本编号为一类;其余的分为一类。这与 IRIS 通常的三类的分类结果中的第一类保持了一致,只是把后两类合并成了一类。

表 1 几种模糊聚类算法的聚类正确度

算法	聚类正确度/%				
	类 1	类 2	类 3	类 4	类 5
C-均值聚类	90	88	100	90	92
模糊 C-均值聚类	90	90	100	90	94
核聚类	90	89	100	92	94
修正核聚类	98	89	95	91	90

表 2 Wine 数据集的聚类结果(每类错聚类个数)

样本集	类别	错聚类个数		
		FCM	KFCM	MKFCM
Wine(178)	1	1	5	5
	2	34	16	10
	3	21	27	16
IRIS(150)	1	8	7	0
	2	8	7	0

上述例子都是针对完整数据集的实验。为了说明修正核函数聚类算法的健壮性,实验采用了 Wisconsin Breast Cancer 数据^[17],共有 10 个特征值及 1 个输出值,由于第 1 个特征值仅为病患的 ID,不影响最后的诊断结果,实验中采用了后 9 个特征值。通过乳腺肿瘤患者的肿瘤块的大小等 9 个特征值,判断此肿瘤为良性或恶性肿瘤。数据集中共有 699 组采样数据。数据集有 16 个特征量的第六个数据属性值缺失。一般做法是把这些异常数据剔除,本文选择让这 16 个数据留在数据集中并设置缺失值为 12(正常取值 1~10)。这样设置后的 16 个数据有明显的误差成分。实验设定预聚类数为 10,其余参数同上。运行 5 次取其平均结果。

表 3 带缺失值的 16 个特征向量

编号	9 个特征									标记	行号
1057013	8	4	5	1	2	12	7	3	1	4	24
1096800	6	6	6	9	6	12	7	8	1	2	41
1183246	1	1	1	1	1	12	2	1	1	2	140
1184840	1	1	3	1	2	12	2	1	1	2	146
1193683	1	1	2	1	2	12	1	1	1	2	159
1197510	5	1	1	1	2	12	3	1	1	2	165
1241232	3	1	4	1	2	12	3	1	1	2	236
169356	3	1	1	1	2	12	3	1	1	2	250
432809	3	1	3	1	2	12	2	1	1	2	276
563649	8	8	8	1	2	12	6	10	1	4	293
606140	1	1	1	1	2	12	2	1	1	2	295
61634	5	4	3	1	2	12	2	3	1	2	298
704168	4	6	5	6	7	12	4	9	1	2	316
733639	3	1	1	1	2	12	3	1	1	2	322
1238464	1	1	1	1	1	12	2	1	1	2	412
1057067	1	1	1	1	1	12	1	1	1	2	618

表 3 中,4 标记恶性,2 标记良性,行号表示样本在原始 wisc9-699^[17] 数据集中的行号。

在表 4 中,类 1 为良性,类 2 为恶性。可以看到,预聚类数的大小设置对结果影响不大,通常取预聚类数目为最优聚类数的 2~3 倍。同时,在不同的预聚类数下,对于缺失的 16 个向量,属于恶性标记的样本全部被正确分配到了类 2 中。在表 5 中,8 个正常样本被划分成恶性,同样 25 个恶性样本被划分成良性,总的错分率 $33/699 = 0.0472 \approx 4.7\%$ 。

表4 不同预聚类数下的结果

预聚类数	不在类1(458)数	不在类2(241)数
5	6	31
10	9	25
15	8	28

表5 wis9-699 的误差,预聚类数 10

数量	类	正确聚类个数	错聚类个数
458	1	474	8
241	2	225	25

5 结语

通过给出数据的紧致性和分离性的核度量,引入一种新的聚类有效性指标 KKW,利用该指标可以得到数据的最优聚类数目,同时还提出了一种修正的核函数聚类算法(MKFCM)。一般的聚类算法需要事先给定聚类数,而把 KKW 指标用于 MKFCM 算法可以解决未知聚类数目情况下的数据聚类问题。仿真实验表明,MKFCM 的错分率低,对于复杂的 wis9-699 数据集错分率为 4.72%,该算法是有效的。

参考文献:

- [1] AMARI S, WU S. Improving support vector machine classifiers by modifying kernel functions[J]. Neural Networks, 1999, 12(6): 783–789.
- [2] 李红英, 钟波. 支持向量分类机的修正函数[J]. 计算机工程与应用, 2009, 45(24): 53–55.
- [3] 潘庆丰, 陈水利, 陈国龙. 基于核函数的模糊 C 均值聚类算法[J]. 集美大学学报, 2006, 11(4): 369–373.
- [4] FILIPPONE M, CAMASTRA F, MASULLI F, et al. A survey of kernel and spectrum methods for clustering[J]. Pattern Recognition, 2008, 41(1): 176–190.
- [5] LOONEY C G. Fuzzy connectivity clustering with radial basis kernel functions[J]. Fuzzy Sets and Systems, 2009, 160(13): 1868–1885.
- [6] XU R, WUNSCH D. Survey of clustering algorithms[J]. Neural Networks, 2005, 14(3): 645–678.
- [7] LEE M, PEDRYCZ W. The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objections having mixing features[J]. Fuzzy Sets and Systems, 2009, 24(16): 3590–3600.
- [8] WU K L, YANG M S. A Cluster Validity index for fuzzy clustering [J]. Pattern Recognition Letters, 2005, 26(9): 1275–1291.
- [9] BAZDEK J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics, 1974, 3(3): 58–73.
- [10] BAZDEK J C. Numerical Taxonomy with fuzzy sets[J]. Journal of Mathematical Biology, 1974, 1(1): 57–71.
- [11] CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273–297.
- [12] LI X X, BENI G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841–847.
- [13] LOONEY C G. Interactive clustering and merging with a new fuzzy expected value[J]. Pattern Recognition, 2002, 35(11): 2413–2423.
- [14] LIN C T, LEE C S G. Neural fuzzy systems[M]. New Jersey: Prentice-Hall, 1995.
- [15] BILLAUDEL P, DEVILLEZ A, LECOLIER G V. Performance evaluation of fuzzy classification methods designed for real time application[J]. International Journal of Approximate Reasoning, 1999, 20(1): 1–20.
- [16] KIM M, RAMAKRISHNA R S. New indices for cluster validity assessment[J]. Pattern Recognition Letters, 2005, 26(15): 2353–2363.
- [17] MANGASARIAN O L, WOLBERG W H. Cancer diagnosis via linear programming[J]. SIAM News, 1990, 23(5): 1–18.

(上接第 1921 页)

据具体的消费者特征来实施。图 2 是对比划、试穿行为规则的展开,显示了 4 条规则。例如:规则 3 表示如果客户的年龄在 24~35 岁,体型匀称,而服饰比较淡雅,能交谈,但行动比较随便,则客户的行为通常是比划或试穿,不购买。比划或试穿这说明客户有需求意向,但最后未买。原因可能是价格问题,也可能是一些服装特征如尺码、款式风格、细节部件或者颜色等不符合消费者需求,如果是尺码问题回旋余地小,但其他问题应该可能得到解决。

从数据挖掘的结果来看,对于该款服装的分类应该从性别、年龄着手,然后是打扮、言谈、脸谱,这说明厂商的分类大范围内是恰当的,而销售商还需要进一步细分才能提高销售效率。

4 结语

建立这种消费行为与客户外表特征印象关联模型的好处是为营销人员提供一些经验规则,以指导销售人员在有限的时间内把握客户,把 b 型客户尽量转化为 c 型客户,从而创造更大的销售量,供应链的上游成员也可因此获得更大的订单而获利。从另一个角度来说,生产商可根据消费群体的喜爱特征、群体密度来开发有针对性的产品,从而帮助实现小批量生产、个性化地开展快速营销活动。决策树算法的数据挖掘技术,计算速度快,实现起来比较容易,而且现在很多的数据库厂商的产品中都提供了这种功能,容易为一般操作人员使用。要嵌入到自己的小商业系统中,则需自己建立挖掘模型。未来的工作是将服装按性别、年龄、款型分类,然后进行实地跟踪和观察,集中进行规则挖掘,构成服装消费者的外表特征

印象与消费行为关联规则库,指导销售人员的营销。

参考文献:

- [1] 许多项. 网络数据库营销[J]. 商业研究, 2002(18): 121–123.
- [2] 唐晓宇. 个性化消费需求下的网络数据库营销的竞争优势[J]. 商业研究, 2002(4): 94–95.
- [3] 欧阳钟辉, 王欢. 客户关系管理与数据库营销体系[J]. 统计与决策, 2008(18): 165–167.
- [4] LI S T, SHUE L Y, LEE S F. Business intelligence approach to supporting strategy-making of ISP service management[J]. Expert System with Application, 2008, 35(3): 739–754.
- [5] APTE C, WEISS S M. Data mining with decision trees and decision rules[J]. Future Generation Computer Systems, 1997, 13(2): 197–210.
- [6] 何田中, 程从从. 基于 Rough 集的规则抽取技术[J]. 南昌大学学报:工科版, 2007, 29(1): 91–94.
- [7] 罗后平. 数据挖掘在市场营销中的应用[J]. 商业研究, 2003(23): 143–14.
- [8] WEN W. A knowledge-based intelligent electronic commerce system for selling agricultural products[J]. Computers and Electronics in Agriculture, 2007, 57(1): 33–46.
- [9] 张红霞, 黄建军. 消费者个人特征对其超市购买频率的影响[J]. 商业研究, 2005(13): 35–41.
- [10] 王国顺, 权明富, 李小文. 基于客户消费行为细分的营销决策分析[J]. 南开管理评论, 2005, 8(1): 52–56.
- [11] SAFAVIAN S R, LANDGREBE D. A survey of decision tree classifier methodology[J]. IEEE Transactions on System, Man and Cybernetics, 1998, 22(5/6): 660–674.
- [12] 李强. 创建决策树算法的比较研究——ID3, C4.5, C5.0 算法的比较[J]. 甘肃科学学报, 2006, 18(4): 84.
- [13] 黄梯云. 智能决策支持系统[M]. 北京: 电子工业出版社, 2001.