基于 K-均值聚类的无监督的特征选择方法*

张　莉, 孙　钢, 郭　军
(北京邮电大学 信息工程学院, 北京 100876)

摘　要: 模式识别方法首先要解决的一个问题就是特征选择,目前许多方法考虑了有监督学习的特征选择问题,对无监督学习的特征选择问题却涉及得很少。依据特征对分类结果的影响和特征之间相关性分析两个方面提出了一种基于 K-均值聚类方法的特征选择算法,用于无监督学习的特征选择问题。

关键词: 特征选择; 相关性分析; 无监督学习; 聚类
中图法分类号: TP391.4　　文献标识码: A　　文章编号: 1001-3695(2005)03-0023-02

Unsupervised Feature Selection Method Based on K-means Clustering

ZHANG Li, SUN Gang, GUO Jun
(*School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing 100876, China*)

Abstract: The first problem need to be solved in pattern recognition method is feature selection. Now many methods think more about supervised feature selection problem, but involve little about unsupervised feature selection problem. In this paper, a feature selection algorithm based on K-means clustering method is proposed involving classification capabilities of feature vectors and correlation analysis between two features. This method can be used in unsupervised feature selection problem.

Key words: Feature Selection; Correlation Analysis; Unsupervised Learning; Clustering

1 引言

模式识别的主要任务是利用从样本中提取的特征将样本划分为相应的模式类别,特征提取与选择是模式识别中的关键技术之一。一般情况下,只有在特征向量中包含了足够的类别信息,才能通过分类器实现正确分类,而特征中是否包含足够的类别信息却很难确定。为了提高识别率,总是最大限度地提取特征信息,结果不仅使特征维数增大,而且其中可能存在较大的相关性和冗余,因而选择合适的特征来描述模式对模式识别的精度、需要的训练时间和需要的实例等许多方面都影响很大,并且对分类器的构造也起着非常重要的作用。目前已有不少文献中提出了有监督学习的特征选择算法[1~4],但对于无监督学习的特征选择问题却涉及较少。无监督学习的特征选择问题就是依据一定的判断准则,选择一个特征子集能够最好地覆盖数据的自然分类。目前的方法有基于遗传算法的特征选择方法[5]、基于模式相似性判断的特征选择方法[6]和信息增益的特征选择方法[7],这几种方法没有考虑特征之间的相关性和特征对分类的影响。文献[8]提出了一种无监督的特征选择方法,基本思想是:首先用竞争学习算法对样本进行分类,确定分类数;然后将原始特征集划分成多个特征子集,在每一个特征子集计算判断函数 $J=\mathrm{trace}((\sum_C+\sum_S)^{-1}\sum_S)$(其中 $\sum_C$, $\sum_S$ 分别表示类内平均离散度和类间平均距离)的值,选择使判断函数值最大的特征子集,从而确定相应的候选特征;最后计算候选特征和已选择的特征之间的相关系数,若相关系数大于 0.75则放弃候选特征。但是由于特征数或特征不同,不同的特征子集对应的自然分类可能也不同,因而对不同的特征子集使用相同的分类结果,不能有效地描述特征对样本自然分类的影响。本文依据特征对分类结果的影响和特征之间相关分析两个方面提出了一种基于 K-均值聚类的特征选择方法,用于无监督学习的特征选择问题。其基本思想是对每一个特征子集利用 K-均值聚类算法确定其最佳分类数,然后以 DB Index准则设定一个判断函数用于特征选择,最后从选择的特征子集中删除掉相关性较大的特征之一。

收稿日期: 2004-04-14　修返日期: 2004-06-18
基金项目: 教育部跨世纪人才基金重点科研项目(02029)

2 相关的背景知识

2.1 聚类有效性的判断规则

类内离散度和类间距离常被用来判断聚类的有效性,DB Index准则同时使用了类间距离和类内离散度,因而在本文中采用 DB Index准则[1]作为分类有效性的判断准则。DB Index准则基本内容如下:

(1)类内平均离散度 $S_i=\frac{1}{|C_i|}\sum_{X\in C_i}\|X-Z_i\|$　　(1)

其中, Z_i 是 C_i 类的类中心; $|C_i|$ 表示 C_i 类样本数。

(2)类间距离　$d_{ij}=\|Z_i-Z_j\|$　　(2)

即用两个类中心的距离表示类间距离。

(3) DB Index　$DB_k=\frac{1}{k}\sum_{i=1}^{k}R_i$　　(3)

其中 $R_i=\max_{j=1,\ldots,k,j\neq i}\frac{S_i+S_j}{d_{ij}}$, k 是分类数目。

DB Index准则是 DB_k 的值越小,说明分类的效果越好。

2.2 特征之间的相关性分析

本文用式(4)计算两个特征之间的相关系数。相关系数 ρ_{xy} 的绝对值大小表示特征 x, y 相关程度的高低, ρ_{xy} 绝对值越

大，表示相关程度越高。

$$\rho_{ij}=\frac{\sum_{p=1}^{n}(x_{pi}-Z_i)(x_{pj}-Z_j)}{\sqrt{\sum_{p=1}^{n}(x_{pi}-Z_i)^2\sum_{p=1}^{n}(x_{pj}-Z_j)^2}} \quad (4)$$

3 特征选择算法

3.1 聚类数的确定

对每一个特征子集 F_i 我们利用 K-均值聚类算法进行对样本进行聚类并确定对应的聚类数 k_i，使用 DB Index准则作为聚类有效性判断。给定一个数据集 X，在没有给定任何样本分布信息的情况下进行聚类，我们采用迭代的方法。一般情况下，最佳的聚类数不会超过 $k_{max}=\sqrt{n}$[9]。因而迭代算法可以在 $k_{min}=2$ 到$\sqrt{n}$之间进行，并且我们可以根据具体的应用设定一个远小于$\sqrt{n}$的 k_{max}值，聚类数 k_i的确定过程如下：

(1)初始化，$C=2$ $DB^*=\infty$，$k_i=1$。其中，C为类的个数迭代变量，k_i表示最佳的分类个数，DB^*表示最小的 DB值。

(2)利用 K-均值聚类算法对样本进行聚类，我们建立如式(5)所示的判断函数，当 $d_j(i)\leqslant\alpha$ 时（α是设定的门限），聚类结束，并且 $DB_c=DB_c(i)$。

$$d_j(i)=\frac{|DB_c(i+1)-DB_c(i)|}{DB_c(i)} \quad (5)$$

其中，$DB_c(i)$表示聚类数为 C的第 i次聚类 DB的值。

(3)若 $DB^*<DB_c$，则 $DB^*=DB_c$ $k_i=C$。

(4)$C=C+1$，若 $C<k_{max}$则转(2)，否则聚类结束。k_i即是第 i个特征子集对应的最佳分类数。

3.2 选择特征子集的判断规则

两个特征子集 F_i，F_j（$i=1\cdots t$，$j=1\cdots t$，$i\neq j$，t是特征子集的个数）对应的特征不是完全相同的，所以对于不同的特征子集 F_i，F_j求得的 DB_{κ_i}，DB_{κ_j}的值没有直接的可比性，因而我们需要将判断规则进行标准化处理。假设 F_i 对应的分类结果 C_i，则判断函数为

$$crit(F_i,C_i)=DB_{\kappa_i} \quad (6)$$

在 F_i特征子集中使用分类结果 C_i，求得相应 DB的值，则 $crit(F_i,C_i)=DB$，然后定义一个标准的判断函数如式(7)所示，特征子集的选择就是要选择使式(7)最小的 F_i。

$$normalizedcrit(F_i)=\frac{1}{t}\sum_{p=1}^{t}crit(F_i,C_i) \quad (7)$$

3.3 基于 K-均值聚类方法的无监督的特征选择算法

在文献[10,11]中提出选择最好的特征子集比选择最好的特征组成特征子集更好，因而在算法中我们利用序贯删除法进行特征子集的搜索。设 F是原始特征集，特征维数 m，令 $t=m$，$count=1$ $normal=0$，其中 t记录特征子集的个数，$count$记录算法执行次数，$normal$保存前一次选择的最佳特征子集的 normalizedcrit的值。算法基本步骤如下：

(1)从 F中依次删除一个特征 x_i，得到 t个特征子集 F_i，$i=1\cdots t$，对这些特征子集分别采用 3.1节中的方法求其对应的最佳分类数 k_i。

(2)采用 3.2节中描述的选择特征子集的判断规则，选择使式(7)最小的 F_i，$t=t-1$，$F=F_i$。

(3)若 $|normalizedcrit(F_i)-normal|>\beta$（$\beta$事先设定的门限）并且 $count\leqslant m$，则 $normal=normalizedcrit(F_i)$，$count=count+1$，转(1)

(4)对选择的特征子集 F_i利用式(4)进行特征相关性分析，若两个特征的相关系数大于 γ（γ为门限），则删除其中的一个特征。

4 实验结果

对于有监督学习情况，特征选择算法的有效性可以通过分类的准确度来评估，但对无监督学习特征选择算法的有效性的评估不能采用这种方法。我们在验证算法时进行了两个实验，首先选择两个维数较少的人工数据集 Wine，Pima Diabetes进行第一个实验（表 1）。这几个人工数据集已知分类数和每一个样本所属类别，因为这两个数据集的特征维数较少，我们在实验结果中给出了全部特征重要性的降序排序，并列出了采用 ReliefF[10]算法得到的特征顺序（表 2）。图 1描述了利用本文的算法和 ReliefF算法选择的特征进行分类的错误率。然后我们采用由哥伦比亚大学完成数据预处理的 KDD Cup 1999 Data中的网络入侵检测的数据进行第二个实验。该数据集提供了从一个模拟的局域网上采集来的九个星期的网络连接数据，数据集中的每条记录包含了 41维特征，并标注了每条记录所属类别（Normal，Dos，U2R，Probing，R2L）。我们从训练集中抽取了 16 645条记录用于特征选择，在实验结果中给出了完成相关性分析后的前 24维特征（表 2）。实验时取 $\alpha=0.0001$，$\beta=0.0001$，$\gamma=0.75$，并用 BP神经网络和 SVM 两个分类器对测试数据用选择的属性进行了测试（表 3）。

表 1 数据集基本信息

数据集	数据类型	特征维数	样本数	分类数
Wine	Continuous	13	178	3
Pima-Diabetes	Continuous	8	768	2
KDD Cup Data	Continuous+Nominal	41	16 645	5

表 2 本文的算法和 ReliefF算法的特征选择结果

数据集	特征重要性的降序序列（本文方法）	特征重要性的降序序列（ReliefF）
Wine	6,7,12,9,11,10,5,13,1,4,3,8,2	6,9,1,11,5,7,10,4,12,2,13,3,8
Pima Diabetes	8,4,3,1,2,6,7,5	8,1,2,5,6,4,7,3
KDD Cup Data	6,5,1,34,33,36,32,8,27,29,28,30,26,38,39,35,13,24,23,11,3,10,12,4	6,26,12,9,13,27,5,23,31,17,39,21,29,33,20,14,34,37,15,35,36,18,28,22

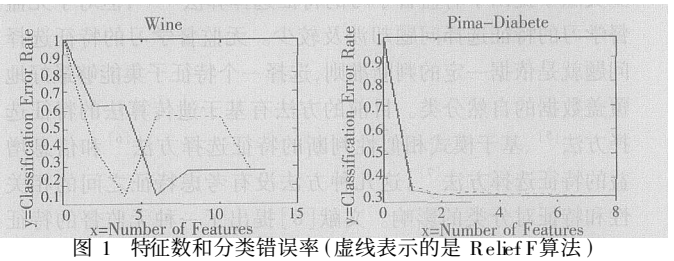

图 1 特征数和分类错误率（虚线表示的是 ReliefF算法）

表 3 入侵检测测试数据集实验结果

分类器	ReliefF算法	本文算法
BP Network	0.1993	0.1017
SVM	0.1020	0.056

5 结论

特征选择是模式识别方法中的难点之一，特别是无监督学习的特征选择问题。本文从特征对分类的影响和特征相关性分析两方面出发，提出了一种基于 K-均值的无 （下转第 42页）

m ing[J]. Artificial Intelligence 1995, 72(1-2): 81-138.
[25] Craig Boutilier Richard Dearden Using Abstractions for Decision-theoretic Planning with Time Constraints[C]. Washington: United States: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, 1994: 1016-1022.
[26] Craig Boutilier *et al*. Exploiting Structure in Policy Construction[C]. Montreal: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995: 1104-1111.
[27] Thomas Dean, *et al*. Planning with Deadlines in Stochastic Domains [C]. Proceedings of the 11th National Conference on Artificial Intelligence: AAAI Press, 1993: 574-579.
[28] Reid Simmons, Sven Koenig. Probabilistic Robot Navigation in Partially Observable Environments[C]. Proceedings of the 14th International Joint Conference on Artificial Intelligence: Montreal: IJCAI Press, 1995: 1080-1087.
[29] C Boutilier, T Dean, S Hanks. Decision Theoretic Planning: Structural Assumptions and Computational Leverage[J]. Journal of Artificial Intelligence Research, 2000, 11: 1-94.
[30] Carlos Guestrin, Daphne Koller, Ronald Parr. Solving Factored POMDPs with Linear Value Functions[C]. Washington: The IJCAI-01 Workshop on Planning under Uncertainty and Incomplete Information (PRO-2), Seattle, 2001: 67-75.
[31] Ronald P A Petrick, Fahiem Bacchus. Knowledge-based Approach to Planning with Incomplete Information and Sensing [C]. Proceedings of the 6th International Conference on Artificial Intelligence Planning Systems (AIPS 2002): Malik Ghallab, Joachim Hertzberg, Paolo Traverso (Eds.), AAAI Press, 2002: 212-222.
[32] Omid Madani, Steve Hanks, Anne Condon. On the Undecidability of Probabilistic Planning and Related Stochastic Optimization Problems [J]. Artificial Intelligence, 2003, 147(1-2): 5-34.
[33] Fox M, Long L. PDDL 2.1: An Extension to PDDL for Expressing Temporal Planning Domains[R]. UK: Technical Report, Department of Computer Science, University of Durham, 2001: 1-54.
[34] Bartak Roman. Modelling Planning and Scheduling Problems with Time and Resources[C]. Proceedings of the 21th Workshop of the UK Planning and Scheduling Special Interest Group (PLANSIG), WSEAS Press: Rethymon, 2002: 87-98.
[35] Coddington A, Fox M, Long D. Handling Durative Actions in Classical Planning Frameworks[C]. Edinburgh, UK: Proceedings of PLANSIG 2001, 2001: 44-58.
[36] Srivastava B, Kambhampati S. Scaling up Planning by Teasing Out Resource Scheduling[R]. Technical Report ASU CSE TR 99-005, Arizona State University, 1999.
[37] Smith, Weld. Temporal Planning with Mutual Exclusion Reasoning [C]. Proceedings of the 16th International Joint Conference on Artificial Intelligence, 1999: 326-337.
[38] Fox M, Long D. The Efficient Implementation of the Plan-graph in STAN [J]. Journal of Artificial Intelligence Research, 1999, 10: 87-115.
[39] Minh B Do, Subbarao Kambhampati. Sapa: A Domain-independent Heuristic Metric Temporal Planner[C]. Spain: Proceedings of the 6th European Conference on Planning (ECP-01) Held in Toledo, 2001: 109-120.
[40] M Fox, D Long. PDDL+: An Extension to PDDL2.1 for Modeling Planning Domains with Continuous Time-dependent Effects[C]. Proceedings of the 3rd International NASA Workshop on Planning and Scheduling for Space, 2003: 1-48.
[41] RE Fikes, N Nilsson. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving[J]. Artificial Intelligence, 1971, 5(2): 189-208.
[42] Jonathan Lever, Barry Richards. parcPLAN: A Planning Architecture with Parallel Actions, Resources and Constraints[C]. Proceedings of 9th International Symposium on Methodologies for Intelligent Systems '94, Springer-Verlag, 1994: 213-223.
[43] A El-Kholy, B Richards. Temporal and Resource Reasoning in Planning: The parcPLAN Approach[C]. Proceedings of the 12th European Conference on Artificial Intelligence, ECAI, 1996: 614-618.
[44] J Penberthy, D Weld. Temporal Planning with Continuous Change [C]. Proceedings of the 12th National Conference of the American Association of Artificial Intelligence, AAAI Press/Mit Press, 1994: 1010-1015.

作者简介:

张友红(1978-),女,吉林汪清人,硕士研究生,主要研究方向为智能规划与规划识别;谷文祥(1947-),男,吉林农安人,教授,主要从事智能规划和规划识别、形式语言与自动机理论、模糊群等研究;刘日仙(1980-),女,浙江江山人,硕士研究生,主要研究方向为智能规划与规划识别。

(上接第24页)监督学习的特征选择方法,适合分类数不确定的情况。但算法时间复杂性较高($O(m^4k^2ne)$,m为维数,k为平均聚类数,n为样本个数,e为聚类算法平均循环次数),对特征维数较高样本数多的情况,计算量较大。从3节中可以看出这主要是由于特征子集的判断规则标准化和最佳分类数的确定引起的,今后的工作是寻找合适的判断规则和确定分类数的方法。

参考文献:

[1] Sergios Theodoridis, Konstantinos Koutroumbas. Pattern Recognition (Second Edition)[M]. 北京:机械工业出版社, 2003: 163-205.
[2] Nojun Kwak, *et al*. Input Feature Selection for Classification Problems [J]. IEEE Transaction on Neural Network, 2002, 13: 143-157.
[3] Ming Dong, Ravi Kothari. Feature Subset Selection Using a New Definition of Classifiability [J]. Pattern Recognition Letters, 2003, 24: 1215-1225.
[4] M Dash. Feature Selection via Set Cover[C]. Newport Beach: Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'97), 1997: 165-171.
[5] M Morita, R Sabourin, *et al*. Unsupervised Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Word Recognition [C]. Edinburgh, Scotland: International Conference on Document Analysis and Recognition (ICDAR'03), 2003: 666-671.
[6] Jayanta Basak, Rajat K De, Sankar K Pal. Unsupervised Feature Selection Using a Neuro-fuzzy Approach[J]. Pattern Recognition Letters, 1998, 19(11): 997-1006.
[7] M Dash, H Liu, J Yao. Dimensionality Reduction of Unsupervised Data[C]. Newport Beach: Proc. 9th IEEE Int'l Conf. Tools with Artificial Intelligence, 1997: 532-539.
[8] Nicolas V, L M, J G Postaire. Unsupervised Color Texture Feature Extraction and Selection for Soccer Image Segmentation[C]. Vancouver, Canada: IEEE International Conference on Image Processing (ICIP'2000), 2000: 800-803.
[9] 于剑,程乾生. 模糊聚类方法中最佳聚类数的搜索范围[J]. 中国科学, 2002, 32(2): 274-280.
[10] Kirak, L A Rendell. A Practical Approach to Feature Selection[C]. The 9th International Conference on Machine Learning: Morgan Kaufmann, 1992: 249-256.
[11] Elashoff J D, *et al*. On the Choice of Variables in Classification Problems with Dichotomous Variables[C]. Biometrika, 1967: 668-770.

作者简介:

张莉(1972-),女,讲师,博士研究生,主要研究方向为智能信息处理和网络安全;孙钢(1972-),男,博士研究生,主要研究方向为计算机网络安全、智能信息处理;郭军(1959-),男,院长,教授,博士生导师,主要研究方向为智能信息处理、网络管理与控制。