一种基于邻域距离的聚类特征选择方法

秦奇伟 梁吉业 钱宇华

(计算智能与中文信息处理教育部重点实验室 太原 030006)

(山西大学计算机与信息技术学院 太原 030006)

摘要 针对高维复杂的符号数据集在聚类中的聚类效果差和计算耗时过大的问题,首先提出了一种基于邻域距离的无监督特征选择算法,然后在选择到的特征子集上进行重新聚类,从而有效提高了聚类结果的精度,降低了聚类计算的计算耗时。实验结果表明,该算法可以找到有效的特征子集,提高数据集的聚类精度,降低面对高维复杂数据集聚类的计算耗时。

关键词 特征选择,聚类计算,邻域距离,属性重要度

中图分类号 TP18 文献标识码 A

Clustering Feature Selection Method Based on Neighborhood Distance

QIN Qi-wei LIANG Ji-ye QIAN Yu-hua

(Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing, Taiyuan 030006, China)

(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract To overcome the limitation of bad results on clustering and time-consuming of existing clustering algorithm to high-dimensional data, we provided an unsupervised feature selection algorithm based on neighborhood distance, then we clustered again on the selected feature subset. The use of the selected feature subset can improve clustering accuracy. The results of the experiment show that the method can find the valid features, and also improve the time-consuming problems in clustering on high-dimensional data.

Keywords Feature selection, Clustering compute, Neighborhood distance, Attribute significance

1 引言

聚类分析^[1]是数据挖掘中的一个重要研究课题,现已被广泛应用于金融欺诈、医疗诊断、图像处理、信息检索和生物信息学等领域^[1-6]。聚类^[1]是将数据对象按照对象间的相似程度进行分类,使得同类中的对象,彼此尽可能相似,而不同类之间的对象彼此尽可能相异。由于其在分类中不需要用训练样本进行学习,因此被称为无监督分类。

然而,在聚类分析的许多应用领域中,随着全球信息化进程的加快,人们通过各种方式得到的原始数据变得越来越庞大和复杂,主要表现为数据的高维和海量。数据的高维引发了数据分析中的“维度效应”问题,使数据变得稀疏、不可聚集^[7]。而且,数据的海量更使得大多数传统的聚类算法在高维空间中失去了作用。因此,当面临高维复杂的数据时,如果能有效利用特征选择得到有效的特征子集,并在特征子集上进行聚类分析,这对于聚类分析的研究是十分必要的。

针对上述问题,很多研究者做了大量的工作以减小甚至消除“维度效应”对数据分析带来的影响^[8,9],目前常用的方法有基于遗传算法的特征选择方法、基于模式相似性判断和信息增益^[10]的特征选择算法,文献^[11]提出了一种基于分级

分类的特征选择方法,这种方法适合于处理标号数据。文献^[12]使用EM聚类的特征选择方法,其基本思想是对每个特征子空间首先确定其最优的分类数,然后选择一个分类器对其进行分类。这一算法适合于解决分类个数不确定的情况,但是复杂度很高,不适合于解决实际问题。

本文通过对原始数据集进行聚类,得到原始数据的一个划分,并将其作为原始数据集的类别信息,提出了一种基于邻域距离的特征选择方法。该算法的主要思想是首先对原始的无类别数据集进行聚类,从而获得数据集的类别标签;然后根据划分结果构成的簇的邻域距离确定数据集中属性的重要度;在此基础上使用启发式搜索的方法进行特征选择。实验结果表明该方法可以得到(使用特征选择所得到的)聚类结果,而且还优于现有的一些无监督特征选择所得到的聚类结果。

本文第2节介绍邻域系统的相关概念;第3节介绍提出的特征选择算法以及本算法的时间复杂度分析;第4节通过实验以及聚类分析指标对所提出的算法进行了验证;最后为本文的结论部分。

2 基本概念

定义1 设 $S=(U,R)$ 为一个信息系统,其中 U 表示对

到稿日期:2011-02-03 返修日期:2011-03-01

秦奇伟(1986—),男,硕士生,主要研究方向为粒计算、数据挖掘、机器学习, E-mail: sxuqinqiwei@126.com; 梁吉业(1962—),男,教授,博士生导师,主要研究方向为粗糙集理论、数据挖掘与知识发现; 钱宇华(1976—),男,博士生,讲师,主要研究方向为粒计算、数据挖掘。

象的非空有限集合,称为论域; R 表示属性的非空有限集合;对任意 $r \in R$ 有 $r:U \rightarrow V_r$,其中 V_r 称为属性 r 的值域;对任意 $r \in R, x \in U$ 有 $f(x, r) \in V_r$,其中 $f(x, r)$ 是一个信息函数,它对论域中对象的每一个属性赋予一个信息值。

定义 2 给定 $S=(U, R)$ 和 $x \in U, x$ 在属性集 $P \subseteq R$ 上的邻域表示为:

$$N_P(x) = \{y | P(x) = P(y), y \in U\}$$

$X \subseteq U, X$ 在属性集 $P \subseteq R$ 上的邻域表示为:

$$N_P(X) = \{y | \forall x \in X, P(x) = P(y), y \in U\}$$

定义 3 给定一个邻域系统 $S=(U, R)$,经聚类后将 U 划分成 l 个类: $C_1, C_2, \dots, C_l, \forall P \subseteq R$,定义类信息 C 关于 P 的邻域表示为:

$$N_P(C) = \{N_P(C_1), N_P(C_2), \dots, N_P(C_l)\}$$

3 特征选择算法

定义 4 给定一个邻域系统 $S=(U, R), X \subseteq U, Y \subseteq U$,定义 X, Y 在属性集 $P \subseteq R$ 上的邻域距离为:

$$D_P(X, Y) = \frac{|(N_P(X) \cup N_P(Y)) - (N_P(X) \cap N_P(Y))|}{|N_P(X) \cup N_P(Y)|}$$

$D(X, Y)$ 可以看作是对象集 X 与对象集 Y 在属性集 $P \subseteq R$ 上的广义距离,对象集之间的距离越大, $D(X, Y)$ 的值也就越大。

定义 5 给定一个邻域系统 $S=(U, R)$,经聚类后将 U 划分成 l 个类: $C_1, C_2, \dots, C_l$,属性 $r \in R$ 的属性重要度为:

$$sig(r) = \left(\sum_{\substack{i=1 \\ j>i}}^l D_r(X_i, X_j) \right) / C_i^2$$

定义 4、定义 5 主要体现了不同特征在不同类之间的区分能力。一个具有聚类性质的特征是指该特征可以使类内对象之间的距离尽可能小,使不同类之间的距离尽可能大。也就是说一个具有聚类性质的特征对于不同的类应该具有较大的区分度。寻找具有聚类性质的特征,可以得到使类内对象间的距离最小、不同类之间的距离最大的特征子集。基于上述分析,本文提出一种基于邻域距离的聚类特征选择算法。该算法主要通过对原始数据集进行聚类来获得数据集的分类标签,在带类别的簇间用邻域距离来度量特征(属性)的重要度,进而选择出具有较大区分度的特征子集。

当对数据集进行特征选择时,计算所有的特征子集对数据集的分类能力,需要测试 $2^m - 1$ 个特征子集(其中 m 为数据集的属性个数),因此对高维海量的数据集,计算耗时将非常大。因此,本文构造了一种启发式搜索算法,该算法以空集为起点,每次加入使当前特征子集重要度最大的属性,直到特征子集的重要度不发生变化。这种启发式搜索算法能够确保重要的属性首先被加入到特征子集,从而不会丢失重要的特征,因此,本算法选出的特征子集作为一个整体能够保持原始数据的分类能力,并且去除了那些与聚类无关的冗余属性和分类能力差的属性。基于邻域聚类的聚类特征选择算法描述见算法 1。

算法 1 一种基于邻域距离的特征选择算法

输入:信息表 $S=(U, R)$,聚类个数 k

输出:特征选择结果

算法步骤如下:

Step1 在数据集上随机选择 K 个初始中心点,并进行 k -modes 聚类,

得到类集 C ;

Step2 $\forall r \in R$,计算重要度 $sig(r)$;

Step3 $\emptyset \rightarrow FS$

Step4 对于任意 $r_i \in R - FS$

计算 $sig(FS \cup \{r_i\})$

Step5 选择 r_i ,其满足

$$\max(sig(FS \cup \{r_i\}))$$

Step6 if $\max(sig(FS \cup \{r_i\})) > 0$

$$FS \cup \{r_i\} \rightarrow FS$$

Go to Step4

Else

Return FS, end

算法 1 的时间复杂度分析如下。

Step1 对原始数据进行聚类,得到类集 C 的时间复杂度为 $O(knmt)$;

Step2 计算所有特征的重要度的时间复杂度为 $O(km^2C_k^2)$;

Step4 的时间复杂度为 $O(km^2C_k^2)$;

Step6 的时间复杂度为 $O(km^2C_k^2m!)$;

通过分析可知,本算法的时间复杂度为 $O(knmt + km^2C_k^2 + O(km^2C_k^2m!))$ 。

针对高维复杂的符号数据集,通过算法 1 可以得到具有聚类性质的有效特征子集,在得到的特征子集上对数据集可进行重新聚类,从而有效降低聚类的计算耗时。

4 实验结果

为了验证本文提出的算法的可行性,下面分别从正确率(accuracy)、类精度(precision)、召回率(recall)^[13] 3 个方面对原始数据集和经过特征选择后的数据集进行聚类分析。

Accuracy(AC), Precision(PE), Recall(RE)分别定义如下:

$$AC = \frac{\sum_{i=1}^k a_i}{n}; PE = \frac{\sum_{i=1}^k \frac{a_i}{a_i + b_i}}{k}; RE = \frac{\sum_{i=1}^k \frac{a_i}{a_i + c_i}}{k}$$

式中, n 表示数据集的对象数, a_i 表示正确分到第 i 个类的对象数, b_i 表示误分到第 i 类的对象数, c_i 表示应该分到第 i 类却没有分到的对象数, k 表示聚类个数。

为了测试本文提出的特征选择算法的有效性,从 UCI 数据集中选取了 3 组分类属性数据集 Chess, Lung-Cancer 和 Soybean,并将经过特征选择后的数据集的聚类分析和原始数据集的聚类分析进行比较。3 组数据描述如表 1 所列。

表 1 数据描述

Data Set	Samples	Attributes	Class
Chess	3196	36	2
Lung-Cancer	32	56	3
Soybean	47	35	4

在每个数据集上,首先运行 k -Modes^[14,15]算法,得到初始的类标签,并记录分类精度等指标;然后运行特征选择算法,记录下特征选择后的特征子集;在得到的特征子集的基础上再次进行聚类分析,并记录新的分类精度。表 2—表 4 给出了在初始数据集上运行了 100 次聚类算法后,通过聚类的正确率、精度、召回率以及标准差 4 个方面验证的算法的有效性。其中, Before FS 表示未经过特征选择的聚类分析结果; After FS 表示在算法 1 得到的特征子集上的聚类分析结果; Mean, Min, Max 和 SD (Standard deviation) 分别表示运行

100次聚类算法后各项评价指标的平均值、最小值、最大值和标准差。

表2 在 Chess 下的性能比较

	Before FS			After FS		
	AC	PE	RE	AC	PE	RE
Mean	0.55018	0.55466	0.68572	0.56277	0.77025	0.9487
Min	0.52222	0.26111	0.5195	0.52222	0.76103	0.82504
Max	0.70651	0.70783	1	0.66051	0.7923	0.9997
SD	0.03645	0.05084	0.09167	0.06281	0.01417	0.07945

表3 在 Lung-Cancer 下的性能比较

	Before FS			After FS		
	AC	PE	RE	AC	PE	RE
Mean	0.72625	0.72956	0.69611	0.74219	0.75265	0.7493
Min	0.71875	0.61677	0.5604	0.71875	0.63462	0.5386
Max	0.8125	0.89655	0.922	0.8437	0.89655	0.9782
SD	0.01546	0.02862	0.07113	0.0408	0.04956	0.08576

表4 在 Soybean 下的性能比较

	Before FS			After FS		
	AC	PE	RE	AC	PE	RE
Mean	0.86702	0.90311	0.91462	0.84617	0.89684	0.9305
Min	0.68085	0.7333	0.70735	0.5744	0.56136	0.84118
Max	0.97872	0.97727	0.985	0.9787	0.9773	0.9853
SD	0.101	0.0656	0.07085	0.1141	0.08795	0.05102

通过分析表2—表4可以看出,数据集 Chess 和 Lung-Cancer 在通过特征选择后的特征子集上的聚类精度得到了显著提高;数据集 Soybean 的召回率 RE 得到了提高。综上,实验结果表明,算法1可以选择到有效的具有聚类性质的特征子集,而且有效提高了数据集的聚类精度。

结束语 本文基于分析数据集中特征的重要度与特征在数据集上的分类能力之间的关系,提出了一种基于邻域距离的特征选择方法,在得到的特征子集上可对数据集进行重新聚类。实验表明,该方法不仅可以选择出具有聚类性质的有效特征子集,而且重新聚类后,数据集的聚类精度得到了显著提高,并降低了针对高维复杂数据集聚类的计算耗时。本文为高维复杂数据集的聚类提供了新的方法和视角。

参 考 文 献

[1] Han Jia-wei, Kamber M. Data Mining Concepts and Techniques [M]. San Francisco: Morgan Kaufmann, 2001
 [2] Brendan J F, Delbert D. Clustering by passing messages between

data points[J]. Science, 2007, 0315(16): 972-976

[3] 张讲社, 梁怡, 徐宗本. 基于视觉系统的聚类算法[J]. 计算机学报, 2001, 24(5): 496-501
 [4] Zhang Jiang-she, Liang Yin-wing. Improved possibilistic c-means clustering algorithms [J]. IEEE Trans on Fuzzy Systems, 2004, 12(2): 2009-217
 [5] 于剑. 论模糊 c 均值算法的模糊指数[J]. 计算机学报, 2003, 26(8): 968-973
 [6] 陈宗海, 文锋, 聂建斌, 等. 基于节点生长 K-均值聚类算法的强化学习方法[J]. 计算机研究与发展, 2006, 43(4): 661-666
 [7] Berchtolds, Bohm C, Keim D A, et al. A cost model for nearest neighbor search in high-dimensional data space [A] // Proceedings of the sixteenth ACM SIGMOD Symposium on Principles of database systems [C]. Tucson, Arizona: ACM Press, 1997
 [8] Ertozi, Stenbachm, Kumarv. Finding clusters of different shapes and densities in noisy high-dimensional data [R]. Minnesota Department of Computer Science, University of Minnesota, 2002
 [9] Ham J H, Lee D D, Saul L K. Learning high-dimensional correspondences from low dimensional manifolds [C] // Proc of ICML Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. Washington, 2003: 34-41
 [10] Dash M, Liu J Y. Dimensionality reduction of unsupervised data [A] // Proc 9th IEEE Int Conf Tool with Artificial Intelligence [C]. 1997: 532-539
 [11] Talavera L. Feature selection as a preprocessing step for hierarchical clustering [A] // Proc 16th Int Conf on Machine Learning [C]. 1999: 389-397
 [12] Dy J G, Brodley C E. Feature subset selection and order identification for unsupervised learning [A] // Proc 17th Int Conf on Machine Learning [C]. 2000: 247-254
 [13] Yang Yi-ming. An evaluation of statistical approaches to categorization [J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88
 [14] Huang Zhe-xue. Clustering large data sets with mixed numeric and categorical values [C] // Proc of PAKDD'97. Singapore: World Scientific, 1997: 21-35
 [15] Huang Zhe-xue. Extensions to the K-means algorithm for clustering large data with categorical values [J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

(上接第 169 页)

出了一种海量规则模式匹配算法。

参 考 文 献

[1] Zhou J J, Yu K M. Tidset-based parallel FP-tree algorithm for the frequent pattern mining problem on PC clusters [J]. Lecture Notes in Computer Science, 2008, 50(3): 18-28
 [2] Aref M M, Tayyri M A. Lana-Match algorithm: A parallel version of the Rete-Match [J]. Parallel Computing, 2008, 24(1): 763-775
 [3] Gaudiot J L, Soha A. Data-driven parallel production systems [J]. IEEE Transactions on Software Engineering, 1990, 16(3): 281-293
 [4] Wolfson O, Ozeri A. Parallel and distributed processing of rules

by data-reduction [J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(3): 523-530

[5] Walzer K, Breddin T G. Matthias relative temporal constraints in the Rete algorithm for complex event detection [C] // Proceedings of the 2nd International Conference on Distributed Event-Based Systems. Rome, ITALY: DEBS 2008: 147-155
 [6] Marsuzawa K. Parallel execution method of production systems with multiple worlds [C] // IEEE Int Workshop Tools Artif Intell Archit Lang Algorithms. SINGAPORE: IEEE, 1989: 339-344
 [7] Forgy C L. Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem [J]. Artificial Intelligence, 1982, 19(1): 17-37
 [8] 张桂刚. 海量规则并行处理研究 [D]. 武汉: 武汉大学, 2009