基于二分 K-均值的 SVM 决策树自适应分类方法*

裘国永，张　娇
（陕西师范大学 计算机科学学院，西安 710062）

摘　要：分析和研究了自适应降维算法在高维数据挖掘中的应用。针对已有数据挖掘算法因维灾难导致的在处理高维数据时准确率和聚类质量都较低的情况，将二分 K-均值聚类和 SVM 决策树算法结合在一起，提出了一种适用于高维数据聚类的自适应方法 BKM-SVMDT。该算法能保证二分 K-均值聚类是在低维数据空间中进行，其结果再反过来帮助 SVM 在高维空间中的执行，这样反复执行以取得较好的分类精度和效率。标准数据集的实验结果证明了该方法的有效性。

关键词：二分 K-均值；支持向量机决策树；降维；自适应算法
中图分类号：TP311.13　　**文献标志码**：A　　**文章编号**：1001-3695(2012)10-3685-03
doi:10.3969/j.issn.1001-3695.2012.10.021

Adaptive SVM decision tree classification algorithm based on bisecting K-means

QIU Guo-yong, ZHANG Jiao
(*School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*)

Abstract: This paper analyzed and researched the applications of adaptive dimension reduction algorithm in high-dimensional data mining. To improve the situation of low accuracy and low clustering quality caused by existing data mining algorithms dealing with high dimensional data, it proposed an adaptively classification algorithm, combining bisecting K-means clustering and support vector machine decision tree, for high dimensional data classification. The BKM-SVMDT algorithm transformed the high dimensional dataset into low dimensional one to ensure data mining in the low-dimensional space, and its results could in turn help SVMDT in high-dimensional space. Adaptively executed the algorithm in order to obtain better classification accuracy and efficiency. Extensive experimental results on standard datasets show the effectiveness of the algorithm.

Key words: bisecting K-means(BKM); SVM decision tree(SVMDT); dimension reduction; adaptive algorithm

0　引言

借助计算机技术进行数据采集和处理是 IT 技术一个极为重要的应用。在数据采集过程中会出现很多高维数据信息，这些信息中包含大量特征。目前数据特征的维度呈越来越高的态势，如文本数据、DNA 实验数据等。这些数据的采集获取相对简单，但当这些特征信息达到一定程度时，就会引发维灾难。维灾难主要是指在数据挖掘过程中，由于维度的增加，空间中的样本个体越来越稀疏，导致了分类等算法由于没有足够的数据样本而无法创建类信息。更重要的是，高维数据致使很多在低维数据空间中行之有效的方法无法应用，从而使数据挖掘的准确率以及质量的下降[1]。为了解决高维数据挖掘的问题，目前主要思路都是将高维数据先进行降维处理，由此出现了两种解决的主要方法，即特征选择和维归约。

特征选择主要是从高维数据的所有特征中找出相关特征的最佳子集，并在该子集中利用各类数据挖掘算法实现分析的目的。特征选择方法的关键在于如何选择最佳的特征子集。目前在特征选择方法方面的研究，主要采用的是过滤模式、封装模式和混合模式。

维归约是处理维灾难的另一种有效方法。维归约的方法主要是将高维数据中两个以上的特征抽取出来后构造出新的特征，从而实现降维的目标。因此维归约也被称做特征抽取。由于其不会造成原始高维数据信息的丢失，被认为是一种解决维灾难的有效途径。对维归约方法的研究，主要是基于维归约技术选择以及利用低维空间中数据挖掘得到的结构构建出高维空间中的类的技术进行[2]。

从总体上看，维归约是利用特征结合的方式，将高维数据变换到低维的空间中，再利用低维的数据挖掘算法进行处理。维归约充分利用了在低维空间中性能和质量表现都较好的数据挖掘算法，能够构建一个简单清晰的数据模型，也降低了数据挖掘算法的时间复杂度和空间复杂度。维归约的方法目前正被越来越多的学者所研究。

近年来，国内外学者在高维数据分类方面也有一定的研究。例如田江等人[3]提出结合高斯过程潜变量模型(GPLVM)和支持向量机(SVM)的阶梯跳跃降维分类框架方法，有效地降低了高维数据集的维数，同时也提高了分类的性能，但是该方法分类速度慢；Niu 等人[4]提出通过引入距离度量标准来修改 SVM 的局部几何结构，从而提高高维数据的分类精度。高维数据分类的分类精度已基本得到解决，但是对于构造分类器的时间需要进一步提高，因此本文提出了保证分类精度的前提

收稿日期：2012-02-25；**修回日期**：2012-04-12　　**基金项目**：陕西省自然科学基金资助项目(2010JM8039)

作者简介：裘国永(1964-)，男，浙江嵊州人，副教授，硕导，博士，主要研究方向为数据挖掘、智能信息处理(qgyqgy@snnu.edu.cn)；张娇(1986-)，女，内蒙古赤峰人，硕士，主要研究方向为数据挖掘.

下,提高分类时间的高维数据分类方法。

1 维归约的一般方法及存在的问题

维灾难直接致使很多在低维空间数据挖掘中表现良好的算法失效,如经典的 K-均值算法等往往只能达到局部最优,无法实现全局最优[5]。目前在维归约方面采用的主要技术是利用线性变换实现从高维数据空间到低维数据空间的映射。主成分分析(PCA)是处理这一问题的常用方法,它通过对高维数据原有特征的线性组合来构造其新特征,而且新构建的特征之间相互正交独立。PCA 属于非监督的维归约方法,被广泛地应用于图像处理、DNA 分析、气象学数据处理等领域。除了 PCA 方法,非监督的维归约方法还包括奇异值分解、随机映射等技术。

利用维归约的数据挖掘算法的具体流程是先利用有非监督或监督的线性变换,将高维数据映射到低维空间,然后在低维空间上再利用数据挖掘算法(如 K-means 聚类等)进行数据分析处理[6]。这些方法的核心是从原始数据的各个特征中构建出新的特征子空间,子空间的维度应大大低于原始空间维度。目前这类方法被广泛地应用于对高维数据进行挖掘和分析的实际过程中。但维归约的方法存在着不少问题,主要是利用线性变换确定了低维子空间后,后继的数据挖掘算法总是基于该子空间进行处理,无法对该空间进行修改,即降维操作与数据挖掘的过程是相互独立的,这类似于特征选择方法中的过滤模式,虽然效率较高,但无法实现最准确的数据分析与处理效果。为了解决这一问题,出现了自适应的维归约方法。这一方法与单纯的维归约方法之间最大的区别在于将降维过程与数据挖掘过程结合到一起,用数据挖掘的过程对低维的子空间进行自适应调整,然后实施挖掘,从而得到最佳的效果。

SVM 是数据分类的强大工具,但传统的 SVM 在解决二次规划问题时速度往往很慢,计算复杂度又较高。目前已有一些改进的 SVM 多分类方法,Mangasarian 等人对传统的 SVM 进行了一些修改,提出了一些较好的 SVM 分类算法,如 proximal SVM(PSVM)[7]、Lagrangian SVM(LSVM)[8]、finite Newton Lagrangian SVM(NLSVM)[9],这些算法都是简单而较快速的算法。其中 NLSVM 算法是用一个等价的分段二次最小函数代替非负约束的最小问题,将 LSVM 中的有约束问题转换为无约束,从而方便使用无约束的牛顿方法。本文选用 NLSVM 来进行对比,以体现本文算法的高效性。

本文选择支持向量机决策树(SVMDT)算法来实现维归约,提出结合二分 K-均值聚类和 SVMDT 的适合于处理高维数据的自适应分类方法,即 BKM-SVMDT 方法。在该方法中,首先用 PCA 将高维数据集变换成低维数据集,然后在低维数据集上执行二分 K-均值聚类来得到样本的类信息,再利用用于表明高低维数据之间的对应关系的指示矩阵 H 生成高维数据的类信息,指导 SVMDT 算法进行分类;得到低维数据集和新的指示矩阵 H 之后,又可以在新的低维数据集上进行二分 K-均值聚类,这个过程可以反复进行下去。这样既能很好地避免了维灾难问题,又能自适应地得到某种形式的收敛结果。本文最后的实验结果证明了 BKM-SVMDT 方法是有效的。

2 基于二分 K-均值的 SVM 决策树高维数据自适应分类方法 BKM-SVMDT

设计一个自适应维归约算法需要解决三个问题:a)降维算法的设计与选择;b)在低维空间所采用的数据挖掘技术;c)如何由低维子空间得到的数据挖掘结果构造出原始高维数据样本的类信息。

2.1 支持向量机决策树算法和二分 K-均值算法

SVM 算法是由 Vapnik 和其他学者提出并发展起来的,主要应用于对小样本、非线性以及高维模式识别等方面。

SVM 算法的核心是找到一个向量 w(超平面系数)来使不同分类之间的距离最大。为了得到这个超平面系数,需要求解一个优化问题,求解方案包括两种:a)利用一个二次的规划器来实现;b)将上述优化问题变换为它的对偶的形式,然后进行优化和求解。

基于传统二叉树的 SVM 算法就是在 SVM 基础理论上发展起来的多分类方法,而传统二叉树的 SVM 分类器具有诸多不足,如性能好坏一定程度上取决于树的结构及对先验知识的依赖、推广性能不好等。本文将利用基于分离度的 SVM 决策树(SVMDT)算法。算法思想是先计算 l 类数据空间中各类的分离度,并将各类分离度进行排序,依次选取分离度最大的类分离出来,构成每个节点的 SVM 分类器,并记录每个节点所用的特征,直至构造 $l-1$ 个 SVM 分类器。

SVMDT 算法流程如下:

输入:$D=\{d_1,d_2,\cdots,d_n\}$ 和簇集 $S=\{C_1,C_2,\cdots,C_l\}$ 及类信息,目标簇数 l。

输出:低维数据集 D' 和 H。

a)在簇集 S 中计算各类的相对分离度 $K_i^D(i=1,2,\cdots,l)$。

b)将相对分离度按降序排列,设 $K_{m1}^D\geqslant K_{m2}^D\geqslant\cdots\geqslant K_{ml}^D$。

c)令计数器 $k=1$。

d)构造子分类器 SVM_k 的训练集 $S_k=\Sigma_1\cup\Sigma_2$,其中 $\Sigma_1=\{(d_{mk},+1)\}$,$\Sigma_2=\{(d,-1)|d\in\{D-\{d_{mk}\}\}$,即把属于相对分离度最高的那个类的 d_{mk} 放入 Σ_1 中。按两类问题构造分类器 SVM_k,并记录 SVM 二分类时用到的特征。

e)调整训练集 D 和计数器 k,其中 $D=D-\{d_{mk}\}$,$k=k+1$。

f)重复 d)和 e),直到构造完第 $l-1$ 个子分类器 SVM_{l-1}。

SVMDT 算法的执行过程需要有类标号,所以不能直接使用在非监督的任务中。数据挖掘算法中的二分均值(BKM)算法可以弥补这一不足。因此将 SVMDT 与 BKM 结合在一起就可应用到非监督的数据挖掘任务中。

二分 K-均值算法是对基本 K-均值数据聚类算法的一种改进与扩充,其主要思想是:假设要将样本数据分为 k 个簇,先用 K-均值算法将所有的数据分为两个簇,然后再从得到的簇中选择一个,继续使用 K-均值算法进行分裂操作,直到获得 k 个簇为止。

二分 K-均值算法的流程如下:

输入:训练数据集 D,分解次数 m,目标簇数 k。

输出:簇集 $S=\{C_1, C_2, \cdots, C_k\}$。

a)初始化簇集 S,它只含一个包含所有样本的簇,并将簇数 k'初始化为 1。

b)从 S 中取出一个最大的簇 C。

c)使用 K-均值聚类算法对簇 C 进行 m 次二分聚类操作,得到 m 对子簇。

d)分别计算这 m 对子簇的总 SSE 的大小,将具有最小总 SSE 的一对子簇添加到 S 中,执行 k' ++操作。

e)如果 $k'=k$,算法结束;否则重复步骤 b) ~ e)。

在二分 K-均值算法中,使用误差平方和来度量聚类的质量,具体的操作是对每一个样本点的误差采用欧氏距离进行计算,最后计算得到误差的平方。二分 K-均值算法没有初始化问题,其每一步操作实际上就是从 m 对子簇中找到误差平方和最小的一对子簇。

因为二分 K-均值算法始终是在低维数据空间中执行,因此在本算法中需要有某种机制由它的聚类结果得到原始数据的类信息。为此要引入一个高低维数据指示矩阵 H,用于表明高低维数据之间的对应关系。设 $D=\{d_1, d_2, \cdots, d_n\}$, $D'=\{d_1', d_2', \cdots, d_m'\}$,这一变换矩阵的构成如下所示:

$$H=(h_{ij})_{n\times m}$$

如果数据 d_j'是由 d_i 降维所得,则 $h_{ij}=1$,否则 $h_{ij}=0$。

在算法的执行过程中,一旦用 SVM 算法挑选出其分类所用的特征,就可以将高维数集 D 变成低维数据集 D';然后让 BKM 在 D'上进行聚类,再借助 H 从 D'的聚类结果构造出原始数据的类信息。

2.2　基于 BKM 的 SVMDT 高维数据自适应分类方法

在自适应维归约数据挖掘技术方法中,都是先利用某种维归约技术得到一个低维数据挖掘空间,再在低维空间上运用数据挖掘算法得到初步的结果。因此在本算法中,要先将高维数据集 D 变换成低维数据集 D',再在低维数据集 D'上进行二分 K-均值聚类算法。首先要有一个变换矩阵 U,它可以通过主成分分析方法得到;然后利用线性变换 $D'=U^{\mathrm{T}}D$ 将高维数据集 D 变换成低维数据集 D',在 D'上执行二分 K-均值聚类算法。

通过上一节的分析,确定了自适应维归约数据挖掘算法中各个关键环节所使用的技术后,下面对基于 BKM 的 SVMDT 高维数据自适应分类方法 BKM-SVMDT 的总体流程进行分析。

BKM-SVMDT 算法流程如下:

输入:训练数据集 D,分解次数 b,循环次数 t,目标簇数 l。

输出:簇集 $S=\{C_1, C_2, \cdots, C_l\}$。

a)在 D 上执行 PCA 算法,获得初始化变换矩阵 U。

b)在 $D'=U^{\mathrm{T}}D$ 上调用 BKM 算法获得指示矩阵 H 和簇集 $\{C_1^*, C_2^*, \cdots, C_l^*\}$。

c)通过指示矩阵 H 由簇集 $\{C_1^*, C_2^*, \cdots, C_l^*\}$ 得到 D 中的簇集 $S=\{C_1, C_2, \cdots, C_l\}$。令计数器 $c=1$。

d)在 D 上执行 SVMDT 算法,利用标记的特征构造出降维后的数据集 D'和指示矩阵 H。

e)在 D'上执行 BKM 得到簇集 $\{C_1^*, C_2^*, \cdots, C_l^*\}$;

f)再由簇集 $\{C_1^*, C_2^*, \cdots, C_l^*\}$ 通过 H 构造出 D 的分类 $\{C_1, C_2, \cdots, C_l\}$。$c=c+1$。

g)重复步骤 d) ~ f),直至收敛或循环次数达到最大。

BKM-SVMDT 算法自适应地在原始数据空间上执行 SVMDT 算法,而在决策子空间上执行 BKM 算法,两者交替进行,直至算法结束。

BKM-SVMDT 算法主要是利用 SVM 决策树分类算法进行分类并实现高维数据到低维数据的转换,再利用 BKM 算法在低维空间得到聚类结果,通过聚类结果再反构出高维数据的类信息。BKM 算法始终是在低维空间中运行。由于在低维空间中利用 BKM 算法得到的结果并不一定是最优解,所以要在原始空间中再次利用 SVM 决策树方法进行新一轮的运行,直至找到最佳的结果。

3　实验结果

实验环境是 Windows XP Professional, MATLAB 7.0, Intel® Pentium® D CPU 2.66/2.67 GHz, 960 MB 内存。为了验证 BKM-SVMDT 算法的有效性,本文将利用不同样本数和维数的数据集进行实验,分别比较 NLSVM 算法、SVM 决策树算法和 BKM-SVMDT 算法的分类精度及运行时间。实验共包括五个数据集,分别是来自于 UCI 数据库的 Gigits、Ionosphere 和 BUPA Liver,以及在文本分类研究中经常使用的 CSTR 和 Log。在实验中选用数据集样本的 2/3 作训练集,样本的 1/3 作测试集,实验所用数据集如表 1 所示。

表 1　实验所用的数据集

数据集	样本数	维数
Gigits	7 494	16
BUPA Liver	345	6
Ionosphere	351	34
CSTR	475	1 000
Log	1 367	200

对于实验结果评价采用的是分类精度指标,即用正确分类的个数除以总个数,得到其算法的分类精度。表 2 是各种不同算法对实验数据集进行处理后的实验结果。

表 2　不同算法的分类精度及运行时间对比

data set($m\times n$)		NLSVM	SVM 决策树	BKM-SVMDT
Gigits 7494×16	train/%	89.5	90.1	97.5
	test/%	89.1	90.0	97.3
	time/s	236.57	242.91	254.87
BUPA Liver 345×6	train/%	75.7	76.1	76.2
	test/%	73.1	74.4	76.0
	time/s	25.54	25.67	29.86
Ionosphere 351×34	train/%	96.1	96.3	98.7
	test/%	95.0	96.1	98.4
	time/s	23.27	25.26	28.89
CSTR 475×1000	train/%	82.3	82.4	85.4
	test/%	82.0	82.1	85.2
	time/s	31.49	32.54	35.93
Log 1367×200	train/%	92.3	92.3	94.7
	test/%	92.1	92.2	94.5
	time/s	68.62	71.53	82.47

由表 2 可以看出,在所选用的标准数据集上,BKM-SVMDT 算法具有较高的分类精度,特别是在高维数据集上,比 NLSVM 算法和 SVM 决策树算法的分类精度有明显提高。对于样本数较多的数据集,BKM-SVMDT 算法的运行时间 (下转第 3709 页)

(上接第 3687 页)较长,但是仍在人们的容忍范围内。对于低维数据集,BKM-SVMDT 算法的分类精度基本上接近于基于 SVM 的其他两种算法。因此相对于 NLSVM 算法和 SVM 决策树法,BKM-SVMDT 算法应用在高维数据集上有更好的效果,具有更强的竞争力。

4 结束语

本文在分析了 SVM 决策树和二分 K-均值算法的基础上,将其相互结合,应用于解决高维空间数据分类的问题,并通过实验证明了其具有较高的分类精度。在该算法中,首先利用 PCA 将高维数据集变换成低维数据集,在低维数据集上执行二分 K-均值聚类方法将数据分成簇,再利用 SVM 决策树算法在高维数据集上分类,利用其结果构造低维数据空间,再在低维空间上进行二分 K-均值聚类,这个过程反复进行下去,不断优化前面的分类结果,直至得到最优结果。该方法有效地解决了高维数据空间产生的维灾难问题,通过降维方法使适用于低维数据集中的数据挖掘算法能够更好地发挥效用。与先前的一些算法比较,BKM-SVMDT 算法改进了它们的一些缺点。对一些标准高维数据集的实验结果也证明了 BKM-SVMDT 算法的有效性。

参考文献:

[1] LI Tao, MA Sheng, OGIHARA M. Document clustering via adaptive subspace iteration[C]//Proc of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press,2004:218-225.

[2] WATANABLE K, AKAHO S, OMACHI S, *et al*. Simultaneous clustering and dimensionality reduction using variational Bayesian mixture model [C]//Proc of the 11th IFCS Biennial Conference and the 33rd Annual Conference of the Gesellschaft für Klassifikation e. V. 2010:81-89.

[3] 田江,顾宏. 高维数据分类方法研究[J]. 系统仿真学报,2009,21(10):2933-2935,2955.

[4] NIU Yan-min, WANG Xu-chu. Improving SVM via local geometric structure for high dimensional data classification[C]//Proc of International Conference on Computer Science for Environmental Engineering and Ecoinformatics. [S. l.]: Springer,2011:299-304.

[5] DUDA R O, HART P E, STORK D G. Pattern classificatio[M]. 2nd ed. New York: John Wiley,2000:35-37.

[6] YE Jie-ping, TAO Xiong. Null space versus orthogonal linear discriminant analysis[C]//Proc of the 23rd International Conference on Machine Learning. New York: ACM Press,2006:80-82.

[7] FUNG G, MANGASARIAN O L. Proximal support vector machine classifiers[C]//Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press,2001:77-86.

[8] MANGASARIAN O L, MUSICANT D R. Lagrangian support vector machines [J]. Journal of Machine Learning Research,2001,1:161-177.

[9] FUNG G, MANGISAIIAN O L. Finite Newton method for lagrangian support vector machine classification[J]. Neurocomputing,2003,55(1-2):39-55.