郭虎升,王文剑,白龙飞.基于向量余弦的支持向量机主动学习策略[J].计算机科学与探索,2014,8(7):868-876.

ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2014/08(07)-0868-09
doi: 10.3778/j.issn.1673-9418.1403056

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

基于向量余弦的支持向量机主动学习策略*

郭虎升[1],王文剑[1,2+],白龙飞[1]
1. 山西大学 计算机与信息技术学院,太原 030006
2. 山西大学 计算智能与中文信息处理教育部重点实验室,太原 030006

Support Vector Machine Active Learning Strategy Based on Vector Cosine*

GUO Husheng[1], WANG Wenjian[1,2+], BAI Longfei[1]
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2. Key Laboratory of Computational Intelligence and Chinese Information Processing, Shanxi University, Taiyuan 030006, China
+ Corresponding author: E-mail: wjwang@sxu.edu.cn

GUO Husheng, WANG Wenjian, BAI Longfei. Support vector machine active learning strategy based on vector cosine. Journal of Frontiers of Computer Science and Technology, 2014, 8(7): 868-876.

Abstract: This paper proposes a support vector machine (SVM) active learning strategy based on vector cosine for the high dimensional dataset to solve the problem that the traditional support vector machine based on active learning can not measure the correlation degree of different samples by Euclidean distance and obtains the low generalization ability, namely COS_SVMactive method. By measuring the information redundancy of training samples based on vector cosine on active learning procedure, several the most valuable samples are selected and need be labeled by experts. In each samples labeling loop, the balance of labeled data is gradually adjusted in order to achieve good generalization performance. The experimental results demonstrate that, compared with common SVM active learning based on random sampling (RS_SVMactive) and SVM active learning based on distance (DIS_SVMactive) methods, the proposed COS_SVMactive method can not only improve classification accuracy, but also reduce the artificial labeling cost.

Key words: support vector machine; active learning; vector cosine; redundancy; balance

---

* The National Natural Science Foundation of China under Grant Nos. 61273291, 60975035 (国家自然科学基金); the Research Project Supported by Scholarship Council of Shanxi Province under Grant No. 2012-008 (山西省回国留学人员科研资助项目); the Graduate Innovation Project of Shanxi Province under Grant No. 20133001 (山西省优秀研究生创新项目).

Received 2014-03, Accepted 2014-05.

CNKI网络优先出版:2014-05-13, http://www.cnki.net/kcms/doi/10.3778/j.issn.1673-9418.1403056.html

摘　要:针对传统基于主动学习的支持向量机(support vector machine,SVM)方法中所采用的欧式距离不能有效衡量高维样本之间的相关程度,导致学习器泛化能力下降的问题,提出了一种基于向量余弦的支持向量机主动学习(SVM active learning based on vector cosine)策略,称为COS_SVMactive方法。该方法通过在主动学习过程中引入向量余弦来度量训练集中样本信息的冗余度,以挑选那些含有重要分类信息的最有价值样本交给专家进行人工标注,并在迭代的样本标注过程中对训练集的平衡度进行逐步调整,使学习器获得更好的泛化性能。实验结果表明,与传统基于随机采样的SVM主动学习方法(SVM active learning based on random sampling,RS_SVMactive)和基于距离的SVM主动学习方法(SVM active learning based on distance,DIS_SVMactive)相比,COS_SVMactive方法不仅可以提高分类精度,而且能够减少专家标记代价。

关键词:支持向量机;主动学习;向量余弦;冗余度;平衡度

文献标志码:A　**中图分类号:**TP18

1　引言

在用数据挖掘方法处理实际生活中遇到的问题时,由于技术和成本的限制,对大规模的数据往往很难获取到足够多的带标记样本。因此,如何利用未标记样本进行无监督或半监督的学习成为当今机器学习领域的研究热点之一。主动学习(active learning,AL)是一种利用未标记样本进行学习的半监督学习方法,其核心在于"主动地"从给定的未标记样本集中挑选出最有价值样本,交由专家进行标记,然后加入训练集,并学习得到分类器,最后通过多次迭代过程改善分类器的性能[1-2]。主动学习方法从功能结构上可分为学习模块和选择模块,前者负责学习器的训练,后者负责从给定的未标记样本集中选出最有价值样本,后者是主动学习方法的关键。按照未标注样本获取方式的不同,可以将现有的主动学习中最有价值样本的选择策略大致分为基于池和基于流的两类。

目前已有学者提出了一些典型的主动学习方法。Tong[1]提出了基于距离的主动学习方法,即选择距离超平面较近的未标记样本作为最有价值样本进行标记。但该方法对于模型选择敏感,特别对于高维数据由于距离的衡量不好控制,效果时好时坏。Seung等人[3]提出了委员会投票(query by committee)方法,该方法把已建立的多个分类器都作为委员会成员,并对未标记样本进行投票,根据投票情况选择最有价值样本交由专家进行标记。该方法虽然效率较高,但存在错误的累积效应。类似采用委员会投票思想对未标记样本的价值进行评判的改进方法还有很多[4-5]。Nguyen等人[6]提出了将主动学习与聚类相结合的改进方法,即将未标记样本集的先验分布知识加入到训练过程,对训练数据进行聚类,通过聚类结果所提供的有效信息提取最有价值样本交由专家进行标记,获得了比传统方法更好的学习效果。Lughofer[7]提出了混合的主动学习模型,该方法将无监督聚类方法和有监督的增量学习方法有机结合,通过增量学习思想选择最有价值样本,有效减少了专家标注未标记样本的代价。李洋等人[8]将主动学习思想与最近邻方法相结合,提出了一种有指导的主动学习方法用于入侵检测,即选择使用少量高质量的训练样本进行建模,以高效完成检测任务。Wang等人[9]则提出了基于自适应规则的主动学习方法,能够自适应地修改当前抽取最有价值样本的规则,以有效指导主动学习中动态样本训练问题中最有价值样本的选择。

支持向量机(support vector machine,SVM)是美国学者Vapnik[10]提出的一类通用有效的机器学习方法,具有坚实的理论基础和良好的泛化能力,目前已成为机器学习领域的研究热点,并在很多领域,如手写数字识别、人脸图像识别、时间序列预测、非线性函数估计等,得到成功的应用。许多学者在SVM算法的优化方面开展了大量工作[11-13],如何将高效的SVM方法与主动学习思想相结合,以处理半监督的分类问题具有重要的理论意义和应用价值。韩光等人[14]采用一种动态聚类过程来选取最有代表性样本,并

根据专家标记与当前SVM分类结果的差值来调整SVM超平面位置,通过这两种策略来解决障碍物检测问题。目前,基于主动学习的SVM方法已经在文本分类[15]、概率估计[16]等领域得到成功的应用。

当所需要处理的样本维度较高(如在文本分类、图像处理等领域)时,传统基于主动学习的SVM方法采用未标记样本到超平面的欧式距离来衡量样本的重要性,以提取最有价值样本。但当样本维度很高时,任意两个样本之间的欧氏距离都会变得很大,即导致所有未标记样本到近似超平面的距离几乎相等,不能准确反应出样本的重要性,无法对样本的冗余程度进行正确度量,不利于主动学习过程中最有价值样本的选择。

针对传统基于主动学习的SVM方法无法有效处理高维数据的问题,本文提出了一种新的基于向量余弦的SVM主动学习(SVM active learning based on vector cosine,COS_SVMactive)策略。该方法针对高维数据定义了新的置信度度量,即将每个样本都看做一个向量,通过衡量两个样本之间的夹角余弦大小来判断样本间的相关程度和冗余信息。这种衡量样本之间差异的方式几乎不受样本维度的影响,可以将样本之间的差异度控制在一个合理范围之内,从而有助于挑选含有最多分类信息的最有价值样本交由专家进行人工标注。此外,在每次迭代标注过程中,根据所得到的相关程度对训练集的平衡度进行调整,获得了较好的泛化能力和较快的收敛速度。

2 COS_SVMactive方法

针对高维数据的半监督分类问题,本文通过向量夹角余弦定义新的样本置信度度量方法,在学习过程中对某些未标记样本的价值进行度量,然后选取价值最大的未标记样本进行人工标注。如此循环往复不断学习,训练SVM分类器,直到需要人工标记的样本被标记完毕或循环次数达到某一阈值时停止。

2.1 置信度度量方法

通常认为离超平面最近的样本最有可能被错分,但它不一定就是最有价值的样本,在度量样本与超平面距离的同时还要考虑样本信息间的冗余情况。当处理维度较低的数据时,传统的欧式距离能够较为准确地反映样本间的相关程度,但处理维度较高的数据时,这种度量往往不再适用。针对这个问题,本文通过度量两个样本(向量)间夹角的余弦判断样本间的相关程度,进而选择最有价值样本交由专家进行标记。

假设 L 和 U 分别表示已标记样本集与未标记样本集,对每个未标记样本 $\boldsymbol{x}_i \in U$ 的置信度定义如下:

$$c(\boldsymbol{x}_i)=\frac{<\boldsymbol{x}_i,\overline{\boldsymbol{x}_j}>}{\|\boldsymbol{x}_i\|\cdot\|\overline{\boldsymbol{x}_j}\|},\boldsymbol{x}_j\in L \tag{1}$$

其中,$\overline{\boldsymbol{x}_j}$ 表示当前已标记样本的均值向量,即:

$$\overline{\boldsymbol{x}_j}=\frac{1}{m}(\sum_{j=1}^{m}\boldsymbol{x}_j),m=|L| \tag{2}$$

$\|\boldsymbol{x}_i\|$ 和 $\|\overline{\boldsymbol{x}_j}\|$ 分别表示未标记样本 $\boldsymbol{x}_i$ 与已标记样本均值 $\overline{\boldsymbol{x}_j}$ 的模;$<\boldsymbol{x}_i,\overline{\boldsymbol{x}_j}>$ 表示向量 $\boldsymbol{x}_i$ 和 $\overline{\boldsymbol{x}_j}$ 的内积。实际上,$\overline{\boldsymbol{x}_j}$ 表示当前迭代过程中已标记样本的平均值,即将 $\overline{\boldsymbol{x}_j}$ 作为当前已标记样本的虚拟代表,然后通过未标记样本 $\boldsymbol{x}_i$ 与 $\overline{\boldsymbol{x}_j}$ 的向量余弦来度量 $\boldsymbol{x}_i$ 与当前已标记样本集的平均相关程度。余弦值越小,两个样本的相关程度越小,即 $\boldsymbol{x}_i$ 与当前已标记样本集包含的信息差异越大,因而就越有价值。

COS_SVMactive方法仅对每个与超平面距离小于当前分类器间隔的未标记样本(即位于分类间隔内的未标记样本) $\boldsymbol{x}_i$ 计算对应的 $c(\boldsymbol{x}_i)$。这样既可降低计算复杂度,又可避免选到孤立点,从而能保证算法快速收敛。然后按所有样本对应的 $c(\boldsymbol{x}_i)$ 的绝对值升序排列,取前 m 个样本(m 为样本抽取参数)进行人工标记,然后加入训练集。

2.2 基于聚类的训练集平衡度调整策略

在每步迭代后得到的带标记样本集都有可能是非平衡的,即超平面可能与一类样本的中心距离较远,与另一类样本中心距离较近。此时,依照本文所提出的置信度度量选择的样本中,样本中心离超平面近的类中样本个数可能会大于另一类样本,如果不对数据集进行处理,则会使算法的学习能力下降。

为避免出现选择最有价值样本导致的数据不平

衡现象,减小样本不平衡性对于模型泛化性能的影响,本文的COS_SVMactive算法在每次迭代后都会检测样本集的平衡度 B,并采用深层聚类的方法进行平衡度调整,其定义如下:

$$B=\begin{cases}\dfrac{num^+}{num^-}, num^+ \leqslant num^- \\ \dfrac{num^-}{num^+}, num^+ > num^-\end{cases} \tag{3}$$

其中,num^+ 表示正类样本的个数;num^- 表示负类样本的个数。当 B 值不大于 ε(ε 为平衡度调节参数)时,集合是不平衡的,此时对多数类数据进行聚类(聚类个数为少数类样本数);然后仅将与聚类中心最靠近的多数类样本与少数类样本加入训练集,而将多数类数据中的其他样本删去,以此来减小训练集不平衡性对泛化性能带来的负面影响,同时也减小模型对于平衡度阈值参数的依赖性。

2.3 COS_SVMactive算法

在COS_SVMactive算法中,每次均选择位于超平面分类间隔内的未标记样本作为最有价值样本,交由专家进行标记并加入训练集,能更快地使算法收敛到较高的精度,获得更好的泛化性能,否则将降低算法的收敛速度甚至不收敛。本文采用了双停止迭代条件,即:(1)若迭代次数达到预设最大迭代参数,则停止迭代;(2)若此时与当前超平面的距离小于当前最大分类间隔的未标记样本都被专家标记完毕,则停止迭代。这样既在一定程度上保证了算法的收敛性,又可使用户在获得满足自己需要的分类精度的同时,将算法的运行时间保持在可控范围内。

假设已标记样本集为 L,未标记样本集为 U,初始标记集合 $L=\varnothing$,所有未标记样本 $U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$;集合 *Need_label* 用来存放主动学习中的最有价值样本,其初始值为空,且每次迭代前均清空;集合 *Train* 中的元素为所有人工标记过的样本,用于SVM训练,其初始值也为空;集合 *Wrong_label* 中的元素为每次迭代中被分错的未标记样本,初始值为空,且同样在每次迭代前都要清空;集合 *U_near* 中的元素为每次迭代中与当前超平面的距离小于当前最大分类间隔的未标记样本,其初始值为空,在每次迭代前同样都要清空;*max_w* 表示当前超平面的最大分类间隔。

算法1 COS_SVMactive算法

步骤1 初始化。将 U 中样本聚为 k 类,对应类中心为 $c_1,c_2,\cdots,c_k$;然后将 $c_1,c_2,\cdots,c_k$ 交由专家标记,若 $c_1,c_2,\cdots,c_k$ 中同时包含正负类样本,则令 $Train=\{c_1,c_2,\cdots,c_k\}$,否则将原始数据聚为 $k+1$ 类,重复这一过程,直到类中心集合同时包含正负类样本为止,$U=U-Train$。初始聚类样本集为 $X^{(0)}$,第 i 层聚类参数为 k_i,初始活动类集 $Set(active)=\varnothing$,初始静止类集 $Set(static)=\varnothing$。

步骤2 循环执行以下训练过程:

步骤2.1 用集合 *Train* 训练SVM,并对 U 中样本类别进行预测。

步骤2.2 对每个 $\boldsymbol{x}_i(\boldsymbol{x}_i\in U)$,计算其与当前超平面的距离 $d(\boldsymbol{x}_i)$,若 $d(\boldsymbol{x}_i)\leqslant max\_w$,则将 $\boldsymbol{x}_i$ 放入 *U_near* 中。

步骤2.3 如果没有找到与当前超平面的距离小于最大分类间隔的样本,即 $\forall\boldsymbol{x}_i(\boldsymbol{x}_i\in U)$,都有 $d(\boldsymbol{x}_i)>max\_w$,此时 $U\_near=\varnothing$,那么转步骤3。

步骤2.4 对 *U_near* 中的样本按式(1)计算 $c(\boldsymbol{x}_i)$,并按 $c(\boldsymbol{x}_i)$ 的值对 *U_near* 中样本进行升序排列,取前 m 个样本加入 *Need_label* 集合,并将 *Need_label* 中的样本交由专家进行标记。

步骤2.5 将 *Need_label* 中各样本的标记结果与步骤2.1中对应样本的标签进行对比,若二者不同,则将其放入 *Wrong_label* 集合。

步骤2.6 按式(3)计算当前 *Wrong_label* 集对应的平衡度 B,若 $B\leqslant\varepsilon$,则按2.2节中的方法对 *Wrong_label* 集的平衡度进行调整。

步骤2.7 $Train=Wrong\_label\cup Train$,$U=U-Wrong\_label$。若循环次数达到预设最大阈值,则转步骤3;否则,继续循环。

步骤3 算法结束。

3 实验及结果分析

为验证本文COS_SVMactive算法的有效性,在4个标准数据集(见表1)上进行了实验。本文的重点在于设计基于向量余弦的SVM主动学习方法,以提取高维数据中最有价值样本交由专家进行标记,提高

SVM处理高维半监督问题的性能,得到优秀的分类结果。由于标准SVM需要存储和计算规模庞大的核矩阵,其时间复杂度和空间复杂度较高,不适用于规模过大的训练样本。关于标准SVM算法效率的问题已经在文献[11-13,17-20]中进行了具体研究,但其均是以牺牲一定的泛化性能作为代价的,这与本文提高高维复杂数据分类精度和泛化性能的研究目的不符,因此SVM本身算法效率改进不作为本文重点讨论内容,实验中采用的训练集规模都控制在100~5 000之间。SVM方法可采用不同的核函数及参数,关于SVM核及参数的选择在文献[21-24]中进行了详细讨论。本文实验中SVM模型全部采用高斯核,核函数参数取1.0,惩罚参数设置为1 000,迭代次数为10,实验中平衡度阈值参数 ε 取0.5。

Table 1 Data sets of experiment

表1 实验采用的数据集

数据集	训练集的规模	测试集的规模	维度
Splice	5 000	10 875	60
Sonar	133	90	60
Hill_Valley_with_Noise	606	606	100
Hill_Valley_without_Noise	606	606	100

3.1 COS_SVMactive算法有效性验证

在同等规模训练集下,以数据集Hill_Valley_without_Noise为例,将COS_SVMactive算法与经典的两种方法RS_SVMactive(SVM active learning based on random sampling)和DIS_SVMactive(SVM active learning based on distance)进行比较。由于RS_SVMactive方法运行结果不稳定,在训练集规模相同的情况下,取10次的平均值作为对比结果,实验中样本抽取参数 m 取30,测试结果对比见表2。

Table 2 Comparison results of Hill_Valley_without_Noise

表2 Hill_Valley_without_Noise上结果比较

方法	准确率/(%)	运行时间/s
RS_SVMactive	72.61	≈0
DIS_SVMactive	69.97	15.687
COS_SVMactive	96.70	33.078

从表2中可以看出,COS_SVMactive与RS_SVMactive、DIS_SVMactive方法相比,在分类精度上提高了20%以上。DIS_SVMactive方法的运行时间虽然较短,但与COS_SVMactive方法相比,其分类精度不理想。而COS_SVMactive算法需要进行样本集的迭代聚类过程,因此运行时间稍长。RS_SVMactive方法的运行时间极短,但其性能不稳,准确率在59.9%~79.21%之间浮动变化,而COS_SVMactive方法准确率稳定且收敛。

两种方法的准确率波动曲线对比见图1。图中横轴对于RS_SVMactive方法表示训练次数,对于COS_SVMactive算法表示迭代次数。从图1中可以看出,在10次实验过程中,COS_SVMactive方法的准确率均高于传统RS_SVMactive方法,且COS_SVMactive方法测试精度的稳定性随着迭代过程基本稳中有升,而传统RS_SVMactive方法则表现不稳定。这说明COS_SVMactive方法随着算法的进程,能够提取到更多的含有重要信息的最有价值样本来进行标记和训练。

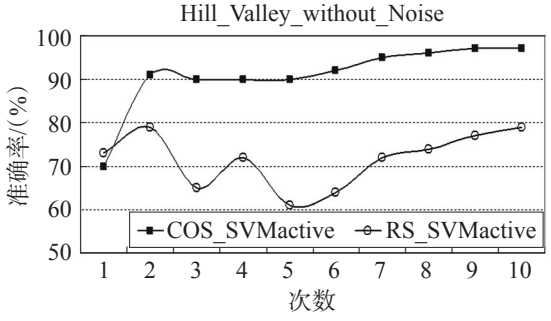

Fig.1 Testing accuracy of RS_SVMactive & COS_SVMactive

图1 RS_SVMactive和COS_SVMactive的准确率

表3是3种方法在Hill_Valley_without_Noise数据集上 m 取不同值时的实验结果。从表3中可以看出,COS_SVMactive方法的 m 值从10变化到30的过程中,精度越来越高,但当 m 值为40时,精度下降,这可能是由于样本过多导致过学习情况的发生;DIS_SVMactive的 m 值从10增加到40时,测试精度也是先升高后降低,同样存在过学习问题,但与COS_SVMactive方法相比低得多;而RS_SVMactive方法虽然随着 m 值增大而增加,但算法精度较低。

Table 3 Testing results of different m on Hill_Valley_without_Noise

表3 Hill_Valley_without_Noise数据集上不同 m 值的结果

方法	m	支持向量个数	准确率/(%)
COS_SVMactive	10	16	94.2
	20	14	95.5
	30	14	96.7
	40	16	92.6
RS_SVMactive	10	14	59.8
	20	19	64.3
	30	25	73.4
	40	32	79.2
DIS_SVMactive	10	17	58.6
	20	19	70.2
	30	18	69.7
	40	21	59.9

表4是COS_SVMactive方法在其他数据集上的实验结果。从表4中可以看出,COS_SVMactive方法在数据集Sonar上得到的分类精度在70%~80%之间,在其他数据集上分类精度都能达到80%或90%以上。此外,从表4中可以看出,除数据集Hill_Valley_with_Noise外,在开始阶段,随着样本抽取参数 m 值的增加,测试的准确率越来越高,但当样本抽取参数达到一定阈值时(如大于30时),精度减小。这是由于开始阶段,随着 m 值增加,每次进入的最有价值样本较多,对于学习器的提高较大,但如果 m 值过大,容易导致学习器产生过学习问题,且迭代过程中测试精度发生振荡,降低了学习器的泛化性能。

为测试本文提出的平衡度阈值参数对测试结果的影响,以及算法1的步骤2.6中对于非平衡数据进行平衡度调整的必要性,在非平衡的 *Need_label* 集合上进行了实验,COS_SVMactive方法通过步骤2.6对训练集进行平衡度调整。以数据集Hill_Valley_without_Noise为例,当 $\varepsilon=0.5$ 时,图2给出了COS_SVMactive方法对训练集平衡度调整前与调整后的测试结果比较。从图2中可以看出,调整训练集平衡度的结果要明显优于调整前,这说明原始数据的不平衡性对结果造成了一定的影响,而采用本文提出的平衡度调整策略有助于在非平衡问题上获得更好的泛化性能。

Table 4 Testing results of COS_SVMactive on other datasets ($\varepsilon=0.5$)

表4 COS_SVMactive在其他数据集上的测试结果($\varepsilon=0.5$)

数据集	m	支持向量个数	准确率/(%)	训练时间/s
Splice	10	19	76.0	0.078
	20	19	76.3	0.094
	30	24	79.8	0.078
	40	25	77.1	0.051
Sonar	10	17	70.0	0.016
	20	8	78.9	0.031
	30	8	78.9	0.063
	40	11	74.4	0.070
Hill_Valley_with_Noise	10	45	90.8	0.734
	20	45	90.1	0.555
	30	39	88.0	0.491
	40	41	89.1	0.909
Hill_Valley_without_Noise	10	16	94.2	0.375
	20	14	95.5	0.302
	30	14	96.7	0.523
	40	16	92.6	0.941

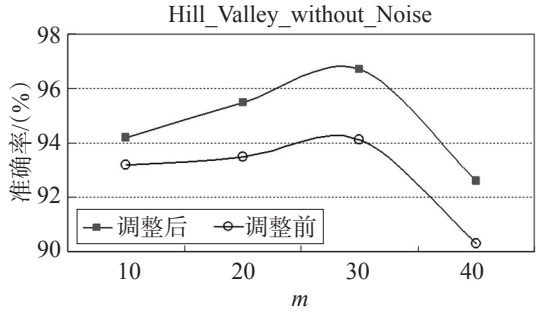

Fig.2 Testing accuracy before and after balance adjusting

图2 平衡度调整前后测试准确率

3.2 COS_SVMactive的收敛性实验

在Sonar数据集上,经过一次算法迭代过程,就达到了表4中的最优实验精度,因此表5~表7分别列出了 m 为10、20、30和40时,COS_SVMactive方法在除Sonar的其他各数据集上迭代10次所得到的分类精度比较。

COS_SVMactive方法在数据集Sonar上第一次迭代过程就达到了最优分类精度,而在其他数据集上

Table 5 Testing accuracy of different m on Splice

表5 m 取不同值时Splice上的分类精度 (%)

m	1	2	3	4	5	6	7	8	9	10
10	68.0	68.0	71.8	70.5	73.5	75.7	75.7	75.7	76.0	76.0
20	66.7	70.2	73.4	73.7	76.3	76.3	76.3	76.3	76.3	76.3
30	66.6	70.3	72.6	74.4	79.8	79.8	79.8	79.8	79.8	79.8
40	67.5	72.1	73.6	77.1	77.1	77.1	77.1	77.1	77.1	77.1

Table 6 Testing accuracy of different m on Hill_Valley_with_Noise

表6 m 取不同值时Hill_Valley_with_Noise上的分类精度 (%)

m	1	2	3	4	5	6	7	8	9	10
10	50.2	59.4	68.7	88.9	88.8	89.9	90.4	90.6	90.8	90.8
20	53.8	78.1	91.1	91.4	91.4	91.4	91.1	91.3	90.1	90.1
30	58.1	85.3	86.0	86.1	87.3	88.0	88.3	88.6	88.8	88.0
40	58.8	86.0	88.0	88.3	88.1	89.8	88.6	88.6	89.0	89.1

Table 7 Testing accuracy of different m on Hill_Valley_without_Noise

表7 m 取不同值时Hill_Valley_without_Noise上的分类精度 (%)

m	1	2	3	4	5	6	7	8	9	10
10	60.0	81.5	90.9	94.1	94.2	94.2	94.2	94.2	94.2	94.2
20	67.2	89.6	91.1	95.9	95.9	95.9	95.2	95.2	95.5	95.5
30	70.0	90.8	90.8	90.8	90.8	92.2	95.9	95.9	96.7	96.7
40	73.9	90.4	90.4	90.3	91.4	91.8	91.8	91.8	93.1	92.6

经过一定的迭代次数后,也取得了较好的收敛效果。这说明COS_SVMactive方法的收敛性较好。从本文只给出了COS_SVMactive方法在不同迭代次数和样本抽取参数下所得到的结果看,COS_SVMactive方法的分类精度最后都收敛到一个较稳定值。

4 结束语

目前,支持向量机已经越来越广泛地应用于各个领域,其与主动学习思想相结合得到的基于主动学习的支持向量机也在大规模半监督或无监督的样本分类问题中以较好的性能获得了广泛关注。针对传统欧式距离不能有效衡量高维样本间相关程度而导致学习器泛化能力低下的问题,本文提出了一种基于向量余弦的支持向量机主动学习策略。该方法在主动学习过程中引入向量余弦来度量训练集中高维样本信息的冗余度,从而挑选那些含有较多分类信息的最有价值样本进行人工标注,并在每次迭代中对训练集的平衡度进行调整,以使学习器获得更好的泛化性能。在分类精度、收敛性、运行时间3个指标上,COS_SVMactive方法都取得了较好的效果。在未来的工作中,结合实际应用问题,将本文提出的基于向量余弦的SVM主动学习方法应用于图像处理、文本分类等高维复杂数据处理领域中。

References:

[1] Tong S. Active learning: theory and application[M]. California: Stanford University Press, 2001: 1-168.

[2] Muslea I, Minton S, Knoblock C A. Active learning with multiple view[J]. Journal of Artificial Intelligence Research, 2006, 27(1): 203-233.

[3] Seung H S, Opper M, Sompolinsky H. Query by committee[C]//Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory (COLT '92). New York, NY,

USA: ACM, 1992: 287-294.

[4] Freund Y, Seung H S, Shamir E, et al. Selective sampling using the query by committee algorithm[J]. Machine Learning, 1997, 28(2/3): 133-168.

[5] Wang Ran, Kwong S, Chen Degang. Inconsistency-based active learning for support vector machines[J]. Pattern Recognition, 2012, 45(10): 3751-3767.

[6] Nguyen H T. Active learning using pre-clustering[C]//Proceedings of the 21st International Conference on Machine Learning (ICML '04), Banff, Canada, 2004. New York, NY, USA: ACM, 2004: 623-630.

[7] Lughofer E. Hybrid active learning for reducing the annotation effort of operators in classification systems[J]. Pattern Recognition, 2012, 45(2): 884-896.

[8] Li Yang, Fang Binxing, Guo Li, et al. Supervised intrusion detection based on active learning and TCM-*K*NN algorithm[J]. Chinese Journal of Computers, 2007, 30(8): 1464-1473.

[9] Wang Zheng, Yan Shuicheng, Zhang Changshui. Active learning with adaptive regularization[J]. Pattern Recognition, 2011, 44(10/11): 2375-2383.

[10] Vapnik V. Statistical learning theory[M]. New York: Wiley Press, l998: 11-23.

[11] Yu H, Yang J, Han J W, et al. Making SVMs scalable to large datasets using hierarchical cluster indexing[J]. Data Mining and Knowledge Discovery, 2005, 11(3): 295-321.

[12] Wang Wenjian, Guo Husheng, Jia Yuanfeng, et al. Granular support vector machine based on mixed measure[J]. Neurocomputing, 2013, 101: 116-128.

[13] Guo Husheng, Wang Wenjian. Dynamical granular support vector regression machine[J]. Journal of Software, 2013, 24 (11): 2535-2547.

[14] Han Guang, Zhao Chunxia, Hu Xuelei. An SVM active learning algorithm and its application in obstacle detection[J]. Journal of Computer Research and Development, 2009, 46(11): 1934-1941.

[15] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2001, 2: 45-66.

[16] Melville P, Yang S M, Saar-Tsechansky M, et al. Active learning for probability estimation using Jensen-Shannon divergence[C]//Proceedings of the 16th European Conference on Machine Learning (ECML '05), Porto, Portugal, 2005. Berlin, Heidelberg: Springer-Verlag, 2005: 268-279.

[17] Irene R L, Carlos S C, Ramon H. Hierarchical linear support vector machine[J]. Pattern Recognition, 2012, 45(12): 4414-4427.

[18] Hao Peiyi, Chiang J H, Tu Yikun. Hierarchically SVM classification based on support vector clustering method and its application to document categorization[J]. Expert Systems with Applications, 2007, 33(3): 627-635.

[19] Bryan C, Narayanan S, Kurt K. Fast support vector machine training and classification on graphics processors[C]//Proceedings of the 25th International Conference on Machine Learning (ICML '08), Helsinki, Finland, 2008. New York, NY, USA: ACM, 2008: 104-111.

[20] Tsang I W, James T K, Cheung P M. Core vector machine: fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005, 6: 363-392.

[21] Wang Wenjian, Xu Zongben, Lu Weizhen, et al. Determination of the spread parameter in the Gaussian kernel for classification and regression[J]. Neurocomputing, 2003, 55(3/4): 643-663.

[22] Zhang Rui, Wang Wenjian. Facilitating the application of support vector machine by using a new kernel[J]. Expert Systems with Applications, 2011, 38(11): 14225-14230.

[23] Shawkat A, Smith-Miles K A. A meta-learning approach to automatic kernel selection for support vector machines[J]. Neural Networks, 2006, 24(1/3): 173-186.

[24] Wu Kuoping, Wang Shengde. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space[J]. Pattern Recognition, 2009, 42(5): 710-717.

附中文参考文献:

[8] 李洋, 方滨兴, 郭莉, 等. 基于主动学习和TCM-*K*NN方法的有指导入侵检测技术[J]. 计算机学报, 2007, 30(8): 1464-1473.

[13] 郭虎升, 王文剑. 动态粒度支持向量回归机[J]. 软件学报, 2013, 24(11): 2535-2547.

[14] 韩光, 赵春霞, 胡雪蕾. 一种新的SVM主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发展, 2009, 46 (11): 1934-1941.

GUO Husheng was born in 1986. He received the Ph.D. degree from School of Computer and Information Technology, Shanxi University in 2014. Now he is a lecturer at School of Computer and Information Technology, Shanxi University, and the member of CCF. His research interests include support vector machine, kernel methods and machine learning, etc.

郭虎升(1986—),男,山西太谷人,2014年于山西大学计算机与信息技术学院获得博士学位,现为山西大学计算机与信息技术学院讲师,CCF会员,主要研究领域为支持向量机,核方法,机器学习等。

WANG Wenjian was born in 1968. She received the Ph.D. degree from Institute for Information and System Science, Xi'an Jiaotong University in 2004. Now she is a professor and Ph.D. supervisor at School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing, Shanxi University, and the senior member of CCF. Her research interests include support vector machine, neural networks, machine learning and environmental computation, etc.

王文剑(1968—),女,山西太原人,2004年于西安交通大学信息与系统科学研究所获得博士学位,现为山西大学计算机与信息技术学院、计算智能与中文信息处理教育部重点实验室教授、博士生导师,CCF高级会员,主要研究领域为支持向量机,神经网络,机器学习,环境计算等。在国内外重要学术刊物和会议上发表学术论文70余篇,主持和参与国家自然科学基金、国家863计划、教育部博士点基金、山西省自然科学基金等项目。

BAI Longfei was born in 1987. He received the M.S. degree from School of Computer and Information Technology, Shanxi University in 2012. His research interests include machine learning and active learning, etc.

白龙飞(1987—),男,山西偏关人,2012年于山西大学计算机与信息技术学院获得硕士学位,主要研究领域为机器学习,主动学习等。