文章编号: 1001-0920 (2015) 08-1356-09 DOI: 10.13195/j.kzyjc.2014.0736

一种基于L2-SVM的多视角核心向量机

黄成泉[1,2], 王士同[1], 蒋亦樟[1], 董爱美[1,3]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 贵州民族大学 工程实训中心，贵阳 550025；3. 齐鲁工业大学 信息学院，济南 250353)

摘 要: 核化一类硬划分SVDD、一/二类L2-SVM、L2支持向量回归和Ranking SVM均已被证明是中心约束最小包含球. 这里将多视角学习引入核化L2-SVM, 提出核化两类多视角L2-SVM (Multi-view L2-SVM), 并证明该核化两类Multi-view L2-SVM亦为中心约束最小包含球, 进而提出一种多视角核心向量机MvCVM. 所提出的Multi-view L2-SVM和MvCVM既考虑了视角之间的差异性, 又考虑了视角之间的关联性, 使得分类器在各个视角上的学习结果趋于一致. 人造多视角数据集和真实多视角数据集的实验均表明了Multi-view L2-SVM和MvCVM方法的有效性.

关键词: 多视角；视角差异性；视角关联性；一致性；核心向量机

中图分类号: TP391.4 **文献标志码:** A

A multi-view core vector machine based on L2-SVM

HUANG Cheng-quan[1,2], *WANG Shi-tong*[1], *JIANG Yi-zhang*[1], *DONG Ai-mei*[1,3]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. Engineering Training Center，Guizhou Minzu University，Guiyang 550025，China；3. School of Information，Qilu University of Technology，Ji'nan 250353, China. Correspondent：HUANG Cheng-quan，E-mail：hcq863@163.com)

Abstract: The kernelized one-class hard-margin SVDD, the kernelized soft-margin one-class and two-class SVMs, the kernelized L2-support vector regression, and the kernelized Ranking SVM can be proved to be the center-constrained minimum enclosing ball(CCMEB) problem. Therefore, a kernelized two-class L2-SVM with multi-view(multi-view L2-SVM) is equivalently formulated as the CCMEB problem, and a classification method named multi-view core vector machine(MvCVM) is proposed. Both the proposed multi-view L2-SVM and MvCVM classifiers can obtain an overall consensus classification result on each view because the differences and the associations between different views are both considered in the two proposed models. An extensive set of experiments on synthetic and real-world multi-view datasets are conducted to demonstrate the effectiveness of the proposed methods.

Keywords: multi-view；differences of different views；associations of different views；consensus；core vector machine

0 引 言

在实际问题中经常会遇到多视角数据. 所谓多视角数据是指数据的来源及处理对象相同, 但观测对象的视角(特征集合)存在一定程度的差异. 例如在进行人口普查时, 可以通过不同的指标(特征)及其组合对某地区的人口收入水平进行分析, 目的在于能够从这些不同的组合数据中分析出较为一致的可靠结果. 多视角学习可分成协作训练类型和协作正则化类型两大类算法[1]. 多视角学习正是将机器学习理论、方法和技术应用于多视角领域而产生的一个热门研究方向, 特别是近几年来, 基于多视角的分类与聚类受到众多学者的关注[2-15]. 上述研究表明, 充分利用多视角数据之间的差异及关联关系能够有效地改善学习性能.

支持向量机(SVM)[16]是一种经典的通过带类标数据来进行训练的机器学习分类算法, 由于较其他分类算法更为简洁有效, SVM算法在机器学习领域得到了广泛的应用. 但是, SVM算法在训练样本数为l时的时间复杂度和空间复杂度分别为$O(l^3)$和$O(l^2)$, 因此该算法在面对大样本数据时效率极低, 从而制约了算法应用范围. 为了对单视角大样本数据进行训练, Tsang等[17]于2005年引入核心集的概念, 提出了

收稿日期: 2014-05-12；**修回日期:** 2014-08-11.

基金项目: 国家自然科学基金项目(61272210, 61202311)；江苏省自然科学基金项目(BK2012552)；贵州省科学技术基金项目(黔科合J字[2013]2136号, 黔科合J字LKM[2013]23).

作者简介: 黄成泉(1976−), 男, 教授, 博士生, 从事模式识别、数据挖掘等研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、数据挖掘、模糊神经网络等研究.

具有渐近线性时间复杂度、而空间复杂度与训练样本数无关的核心向量机(CVM), 并通过最小包含球(MEB)算法, 把单视角SVM算法中求解支持向量的问题转化为求解核心集的MEB问题, 极大地缩短了支持向量的求解时间. 作为CVM工作的延伸, Tsang等[18]于2006年又进一步提出了中心约束最小包含球(CCMEB), CCMEB解决了CVM只能进行一类硬划分SVDD和一/二类L2-SVM而不能用于L2支持向量回归(L2-SVR)和Ranking SVM等机器学习算法. 此后, 基于CCMEB处理大样本数据集的算法不断涌现[19-23]. 然而, 上述以核心向量机为基础的系列算法仍然处于单视角学习框架内, 其在建模过程中因算法本身的单视角特性使其仅能利用单个视角的数据信息, 不具备考虑多个视角数据之间的差异性和关联性, 故得到的分类机只能从当前视角进行学习. 这使得当前视角数据若有重要的信息缺失时, 则受训所得之分类器泛化性能将会变差.

本文将拓展两类核化L2-SVM算法, 使得能够在保持原本的大样本快速学习能力的同时还适用于多视角学习领域, 提出一种多视角L2-SVM和多视角核心向量机(MvCVM). 该方法能够克服现有的单视角方法在多视角数据上泛化性能差的缺陷, 同时能够保持CVM处理大样本数据所具有的优越性能. 实验结果表明: 1) 本文提出的多视角算法充分利用各视角的差异性和关联性, 使得学习效果较单视角算法更趋于一致; 2) 对于大样本多视角数据集, 本文所提出的MvCVM算法在提高训练速度和削减内存需求的同时, 仍然能保持SVM的泛化性能.

1　多视角支持向量机和多视角核心向量机

由上可知, 大样本快速学习问题正得到越来越广泛的关注, 并且诸如CVM的机器学习方法已在该领域取得了一定的成功, 推动了该领域的发展. 描述数据的方式很多, 从量的角度阐述大样本是目前的热点; 而从产生数据的方法及对象研究, 多视角学习也已成为当前研究的热点并受到了广泛的关注. 由于大样本与多视角数据表征不存在矛盾和冲突, 甚至在大量的应用实例中这两者总是同时出现或同时存在, 这使得人们尝试研究出一种既能够具备大样本快速处理功能、又能够进行多视角学习、具备综合考虑各视角特征的算法. 为此, 这里将以L2-SVM算法作为基础算法, 拓展该算法的多视角学习能力, 从而得到具备多视角学习能力同时也具备大样本快速处理性能的MvCVM算法.

1.1　Multi-view L2-SVM 和 MvCVM

给定l个训练样本$\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_l, y_l)\}$, 其中$(\boldsymbol{x}_i, y_i) \in \boldsymbol{S} \times \{1, -1\}$, 学习的目标是学习从$\boldsymbol{S}$到$\{1, -1\}$的映射, 使得此映射在未来数据对象上有好的预测性能. 其中: $\boldsymbol{x}_i \in \boldsymbol{S}$表示数据点, $y_i \in \{1, -1\}$表示数据点$\boldsymbol{x}_i$的标记. 设$\boldsymbol{x}_i^{(t)}$表示数据$\boldsymbol{x}_i$的第t个视角, $\boldsymbol{x}_i = [\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \cdots, \boldsymbol{x}_i^{(v)}]$表示数据点$\boldsymbol{x}_i$由$v$个视角数据构成. 记$\varphi(\boldsymbol{x}_i) = [\varphi_1(\boldsymbol{x}_i^{(1)})^{\mathrm{T}}, \cdots, \varphi_v(\boldsymbol{x}_i^{(v)})^{\mathrm{T}}]^{\mathrm{T}}$, 其中$\varphi_t$为第$t$个视角上的核映射, $t = 1, 2, \cdots, v$. 根据文献[17], 如下二类核化v视角L2-SVM为一MEB问题:

$$\begin{aligned} &\min\ \ \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l}(\xi_i)^2 - \rho;\\ &\text{s.t.}\ \ y_i\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i^{(1)}, \cdots, \boldsymbol{x}_i^{(v)}) \geqslant \rho - \xi_i,\\ &\qquad i = 1, 2, \cdots, l. \end{aligned} \tag{1}$$

其决策函数为

$$f(x) = \mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})) =$$

$$\mathrm{sgn}\Big(\sum_{i=1}^{l}\alpha_i y_i \varphi(\boldsymbol{x}_i^{(1)}, \cdots, \boldsymbol{x}_i^{(v)})^{\mathrm{T}}\varphi(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(v)})\Big) =$$

$$\mathrm{sgn}\Big(\sum_{t=1}^{v}\sum_{i=1}^{l}\alpha_i y_i \varphi_t(\boldsymbol{x}_i^{(t)})^{\mathrm{T}}\varphi_t(\boldsymbol{x}^{(t)})\Big), \tag{2}$$

其中α_i为式(1)的对偶QP问题解.

式(2)表明, 二类核化v视角L2-SVM对新样本点的决策是由新样本点各个视角共同决策的结果, 只有当每个视角的决策一致时才能保证二类核化v视角L2-SVM对新样本点的决策起作用(一致性). 事实上, 二类核化v视角L2-SVM本质上是单视角分类器, 因而并不能在各个视角上取得一致结果.

由于式(1)和(2)不能有效利用各个视角的大量信息来改善分类器学习性能, 为此, 本文基于各视角之间的差异性和关联性建立了如下的多视角分类器Multi-view L2-SVM:

$$\begin{aligned} &\min\ \ \frac{1}{2}\sum_{t=1}^{v}\|\boldsymbol{w}_t\|^2 + \sum_{t=1}^{v}\frac{C_t}{2}\sum_{i=1}^{l}(\xi_i^{(t)})^2 - \rho;\\ &\text{s.t.}\ \ y_i\boldsymbol{w}_t^{\mathrm{T}}\varphi_t(\boldsymbol{x}_i^{(t)}) \geqslant \rho - \sum_{k=1}^{t}\xi_i^{(k)},\\ &\qquad i = 1, 2, \cdots, l,\ t = 1, 2, \cdots, v. \end{aligned} \tag{3}$$

其中: $\boldsymbol{w}_t$、C_t、$\xi_i^{(t)}$分别为视角t上的权重向量、错分惩罚因子、松弛变量.

式(3)既充分考虑了各个视角之间的差异性(式(3)目标函数), 又充分考虑了各个视角之间的关联性(式(3)约束条件). 式(3)目标函数不仅要求整个分类器间隔$\boldsymbol{w}(\boldsymbol{w} = [\boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{w}_2^{\mathrm{T}}, \cdots, \boldsymbol{w}_v^{\mathrm{T}}]^{\mathrm{T}})$最大, 还要求每一个视角上的分类器间隔最大, 并且每一个视角上的错分程度达到最小. 这样通过各个视角之间的相互协作(式(3)), 使得最终的学习效果趋于一致(各个视角

差异最小, 见实验部分).

通过对式(3)引入vl个拉格朗日乘子$\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_{vl}]^{\mathrm{T}}$, 并构造拉格朗日函数进行求解, 可得其对偶为

$$\begin{aligned}&\max\quad -\boldsymbol{\gamma}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\gamma};\\&\text{s.t.}\ \ \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1}=1,\boldsymbol{\gamma}\geqslant\mathbf{0}.\end{aligned}\tag{4}$$

其中

$$\tilde{\boldsymbol{K}}=\frac{1}{2}\begin{bmatrix}\boldsymbol{K}^{(1)}\circ\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}+\dfrac{1}{C_1}\boldsymbol{I} & \cdots & \dfrac{1}{C_1}\boldsymbol{I} & \dfrac{1}{C_1}\boldsymbol{I}\\ \vdots & \vdots & \vdots & \vdots\\ \dfrac{1}{C_1}\boldsymbol{I} & \cdots & \boldsymbol{K}^{(v-1)}\circ\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}+\Big(\sum\limits_{j=1}^{v-1}\dfrac{1}{C_j}\Big)\boldsymbol{I} & \Big(\sum\limits_{j=1}^{v-1}\dfrac{1}{C_j}\Big)\boldsymbol{I}\\ \dfrac{1}{C_1}\boldsymbol{I} & \cdots & \Big(\sum\limits_{j=1}^{v-1}\dfrac{1}{C_j}\Big)\boldsymbol{I} & \boldsymbol{K}^{(v)}\circ\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}+\Big(\sum\limits_{j=1}^{v}\dfrac{1}{C_j}\Big)\boldsymbol{I}\end{bmatrix};$$

$\boldsymbol{K}^{(t)}=[k(\boldsymbol{x}_i^{(t)},\boldsymbol{x}_j^{(t)})]_{l\times l}=[\varphi_t(\boldsymbol{x}_i^{(t)})^{\mathrm{T}}\varphi_t(\boldsymbol{x}_j^{(t)})]_{l\times l}$为第$t$个视角核矩阵, $t=1,2,\cdots,v$; $\boldsymbol{y}=[y_1,\cdots,y_l]^{\mathrm{T}}$; $\mathbf{1}=[1,1,\cdots,1]_{vl\times1}^{\mathrm{T}}$; $\circ$为矩阵的Hadamard乘积运算.

由式(4)可以看出$\mathrm{diag}(\tilde{\boldsymbol{K}})$不恒为常数, 但有如下定理:

定理 1 多视角分类器Multi-view L2-SVM等价于CCMEB问题.

证明 取$\eta=\max(\mathrm{diag}(\tilde{\boldsymbol{K}}))$, $\boldsymbol{\Delta}=-\mathrm{diag}(\tilde{\boldsymbol{K}})+\eta\mathbf{1}$, 则式(4)等价于

$$\begin{aligned}&\max\ \ \boldsymbol{\gamma}^{\mathrm{T}}(\mathrm{diag}(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}-\eta\mathbf{1})-\boldsymbol{\gamma}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\gamma};\\&\text{s.t.}\ \ \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{\gamma}\geqslant\mathbf{0}.\end{aligned}\tag{5}$$

由η的取法有$\boldsymbol{\Delta}=[\delta_1^2,\delta_2^2,\cdots,\delta_{vl}^2]^{\mathrm{T}}\geqslant\mathbf{0}$, 按照文献[18], 式(5)为一CCMEB问题. □

由定理1, 可以利用CCMEB的逼近算法(GCVM)快速获取训练样本核心集, 然后在核心集上求解二次规划(4)并使用下面的决策函数对新样本进行决策. 本文把使用GCVM算法求解多视角分类器Multi-view L2-SVM的方法称为多视角核心向量机(MvCVM).

1.2 决策函数

为了测试新样本点$\boldsymbol{x}=(\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(v)})\in\boldsymbol{R}^d$的类别, 需要通过决策函数进行判断. 容易知道, Multi-view L2-SVM或MvCVM的决策函数为

$$\begin{aligned}&\mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}))=\\&\mathrm{sgn}\Big(\Big(\sum_{i=1}^{l}\gamma_i y_i\varphi_1(\boldsymbol{x}_i^{(1)})^{\mathrm{T}},\cdots,\\&\sum_{i=1}^{l}\gamma_{(v-1)l+i}y_i\varphi_v(\boldsymbol{x}_i^{(v)})^{\mathrm{T}}\Big)\varphi(\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(v)})\Big)=\\&\mathrm{sgn}\Big(\sum_{t=1}^{v}\sum_{i=1}^{l}\gamma_{(t-1)l+i}y_i\varphi_t(\boldsymbol{x}_i^{(t)})^{\mathrm{T}}\varphi_t(\boldsymbol{x}^{(t)})\Big).\end{aligned}\tag{6}$$

其中: $\mathrm{sgn}\Big(\sum\limits_{i=1}^{l}\gamma_{(t-1)l+i}y_i\varphi_t(\boldsymbol{x}_i^{(t)})^{\mathrm{T}}\varphi_t(\boldsymbol{x}^{(t)})\Big)$为第$t$个视角上的决策函数; $t=1,2,\cdots,v$.

式(6)表明, Multi-view L2-SVM和MvCVM对新样本点的决策是由新样本点各个视角共同决策的结果, 只有当每个视角的决策一致时才能保证Multi-view L2-SVM和MvCVM对新样本点的决策起作用(一致性).

由式(4)容易看出, Multi-view L2-SVM为标准QP问题, 为此, 本文仅给出MvCVM算法描述.

1.3 MvCVM算法描述

在采用GCVM算法获取多视角数据样本核心集时, 需要计算高维特征空间中最小包含球的半径和高维特征空间中点到球心的距离. 由于等价的CCMEB问题(5)中含有vl个拉格朗日乘子变量, 在高维特征空间中有vl个点与之对应. 事实上, 从核矩阵$\tilde{\boldsymbol{K}}$可以看出, v视角问题中的任一样本点$\boldsymbol{z}_i=([\boldsymbol{x}_i^{(1)},\cdots,\boldsymbol{x}_i^{(v)}],y_i)$对应于高维特征空间维数增加1维后的$v$个点, 即

$$\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_i)\\ \delta_i\end{bmatrix}=\begin{bmatrix}y_i\varphi_1(\boldsymbol{x}_i^{(1)})\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\\ \sqrt{\dfrac{1}{C_1}}\boldsymbol{e}_i\\ \delta_i\end{bmatrix},$$

$$\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_{i+l})\\ \delta_{i+l}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ y_i\varphi_2(\boldsymbol{x}_i^{(2)})\\ \vdots\\ \mathbf{0}\\ \sqrt{\dfrac{1}{C_1}}\boldsymbol{e}_i+\sqrt{\dfrac{1}{C_2}}\boldsymbol{e}_{i+l}\\ \delta_{i+l}\end{bmatrix},$$

$$\vdots$$

$$\begin{bmatrix} \tilde{\varphi}(\boldsymbol{z}_{i+(v-1)l}) \\ \delta_{i+(v-1)l} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ y_i\varphi_v(\boldsymbol{x}_i^{(v)}) \\ \sum\limits_{t=1}^{v}\sqrt{\dfrac{1}{C_t}}\boldsymbol{e}_{i+(t-1)l} \\ \delta_{i+(v-1)l} \end{bmatrix}.$$

其中: $\tilde{\varphi}$ 为核矩阵 $\tilde{\boldsymbol{K}}$ 对应的特征映射, $\boldsymbol{e}_i$ 为第 i 个分量为1而其余分量均为0的 vl 维列向量, δ_i 的定义见定理1证明.

此外, 为了计算高维特征空间中任何两点之间的距离, 下面给出 v 视角距离定义.

定义 1　设 $\boldsymbol{z}_i = ([\boldsymbol{x}_i^{(1)}, \cdots, \boldsymbol{x}_i^{(v)}], y_i)$ 和 $\boldsymbol{z}_j = ([\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}_j^{(v)}], y_j)$ 为样本数据集中任何两个样本点, φ 为样本空间到高维特征空间的特征映射, $\tilde{\varphi}$ 为核矩阵 $\tilde{\boldsymbol{K}}$ 对应的特征映射, 称

$$\begin{aligned} &d(\varphi(\boldsymbol{z}_i), \varphi(\boldsymbol{z}_j)) = \\ &\max\left(d\left(\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_i)\\ \delta_i\end{bmatrix}, \begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_j)\\ \delta_j\end{bmatrix}\right), \cdots, \right. \\ &\left. d\left(\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_{i+(v-1)l})\\ \delta_{i+(v-1)l}\end{bmatrix}, \begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_{j+(v-1)l})\\ \delta_{j+(v-1)l}\end{bmatrix}\right)\right) \end{aligned}$$

为 v 视角距离, 其中 $d\left(\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_i)\\ \delta_i\end{bmatrix}, \begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_j)\\ \delta_j\end{bmatrix}\right)$ 为高维特征空间中的 l_2 范数距离.

由CCMEB及对偶规划理论知, 问题(3)的中心点参数 $\boldsymbol{c}_{\tilde{\varphi}}$ 和超球半径 $r_{\tilde{\varphi}}$ 可以通过求解二次规划(5)得到最优解 $\boldsymbol{\gamma}$ 后代入下式得到:

$$\begin{aligned} &\boldsymbol{c}_{\tilde{\varphi}} = \sum_{i=1}^{vl}\gamma_i\tilde{\varphi}(\boldsymbol{z}_i), \\ &r_{\tilde{\varphi}} = \sqrt{\boldsymbol{\gamma}^{\mathrm{T}}(\mathrm{diag}(\tilde{\boldsymbol{K}}) + \boldsymbol{\Delta} - \eta\boldsymbol{1}) - \boldsymbol{\gamma}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\gamma}}. \end{aligned} \tag{7}$$

于是高维特征空间维数增加1维后的任一点 $\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_m)\\ \delta_m\end{bmatrix}$ 与中心点 $\begin{bmatrix}\boldsymbol{c}_{\tilde{\varphi}}\\ 0\end{bmatrix}$ 的 l_2 范数距离平方为

$$\begin{aligned} &\|\boldsymbol{c}_{\tilde{\varphi}} - \tilde{\varphi}(\boldsymbol{z}_m)\|^2 + \delta_m^2 = \\ &\tilde{\varphi}(\boldsymbol{z}_m)^{\mathrm{T}}\tilde{\varphi}(\boldsymbol{z}_m) + \delta_m^2 - \\ &2\sum_{i=1}^{vl}\gamma_i\tilde{\varphi}(\boldsymbol{z}_i)^{\mathrm{T}}\tilde{\varphi}(\boldsymbol{z}_m) + \sum_{i,j=1}^{vl}\gamma_i\gamma_j\tilde{\varphi}(\boldsymbol{z}_i)^{\mathrm{T}}\tilde{\varphi}(\boldsymbol{z}_j). \end{aligned} \tag{8}$$

至此, 可以给出MvCVM算法描述.

算法 1　多视角核心向量机(MvCVM)算法.

输入: v 视角数据训练样本 $\boldsymbol{S}$, CCMEB逼近精度阈值 ε 及 η 等参数;

输出: 核心集 $\boldsymbol{X}_k$, 权重向量 $\boldsymbol{\gamma}$.

Step 1: 初始化核心集 $\boldsymbol{X}_0$, 最小包含球中心 $\boldsymbol{c}_{\tilde{\varphi}}^{(0)}$, 半径 $r_{\tilde{\varphi}}^{(0)}$, 迭代次数 $k=0$;

Step 2: 若高维特征空间中所有点都包含在以 $\boldsymbol{c}_{\tilde{\varphi}}^{(k)}$ 为球中心、$(1+\varepsilon)r_{\tilde{\varphi}}^{(k)}$ 为半径的球之内, 则转Step 6, 否则转Step 3;

Step 3: 将高维特征空间中距离球心 $\boldsymbol{c}_{\tilde{\varphi}}^{(k)}$ 最远的点 $\varphi(\boldsymbol{z})$ 选入核心集, 即 $\boldsymbol{X}_k = \boldsymbol{X}_k \bigcup \{\boldsymbol{z}\}$;

Step 4: 对核心集 $\boldsymbol{X}_k$ 求解CCMEB的对偶问题(5), 并用式(7)计算球心 $\boldsymbol{c}_{\tilde{\varphi}}^{(k+1)}$ 和半径 $r_{\tilde{\varphi}}^{(k+1)}$;

Step 5: 置 $k = k+1$ 并转Step 2;

Step 6: 算法停止, 输出核心集 $\boldsymbol{X}_k$ 和权重向量 $\boldsymbol{\gamma}$.

由于MvCVM算法处理的是多视角样本数据, 本文提出的MvCVM算法的实现有别于CCMEB算法. 下面给出MvCVM算法实现的细节描述.

1) 在Step 1中, 虽然可以从训练样本集 $\boldsymbol{S}$ 中任选一点 $\boldsymbol{z}$ 来初始化核心集 $\boldsymbol{X}_0 = \{\boldsymbol{z}\}$, 但好的初始化方法将大大提高算法的性能[23, 25-26]. 本文参照文献[17]方法, 从训练样本集 $\boldsymbol{S}$ 中随机取一点 $\boldsymbol{z}_0$, 然后在高维特征空间中按照定义1找出距离 $\varphi(\boldsymbol{z}_0)$ 最远的一点 $\varphi(\boldsymbol{z}_1)$, 再从高维特征空间中找出距离 $\varphi(\boldsymbol{z}_1)$ 最远的一点 $\varphi(\boldsymbol{z}_2)$, 这样初始化核心集 $\boldsymbol{X}_0 = \{\boldsymbol{z}_1, \boldsymbol{z}_2\}$. 对初始核心集 $\boldsymbol{X}_0$ 求解CCMEB的对偶问题(5), 并由式(7)得到初始球心 $\boldsymbol{c}_{\tilde{\varphi}}^{(0)}$ 和半径 $r_{\tilde{\varphi}}^{(0)}$;

2) 在MvCVM算法Step 2和Step 3中, 需要使用式(8)计算高维特征空间中 v 个点 $\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_m)\\ \delta_m\end{bmatrix}, \cdots, \begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_{m+(v-1)l})\\ \delta_{m+(v-1)l}\end{bmatrix}$ 到球心 $\begin{bmatrix}\boldsymbol{c}_{\tilde{\varphi}}^{(k)}\\ 0\end{bmatrix}$ 的 l_2 范数距离平方. 然而, 第 k 次迭代求解式(8)的时间复杂度为 $O(v^2\cdot|\boldsymbol{X}_k|^2 + v|\boldsymbol{S}|\cdot|\boldsymbol{X}_k|)$, 当样本 $\boldsymbol{S}$ 规模较大时不利于算法求解. 本文算法实现时采用Smola等[27]提出的概率加速方法, 即在训练样本集 $\boldsymbol{S}$ 中随机选取一大小为59的子集 $\boldsymbol{S}_{\mathrm{sub}}$, 在子集 $\boldsymbol{S}_{\mathrm{sub}}$ 所对应的高维特征空间中寻找离球心 $\begin{bmatrix}\boldsymbol{c}_{\tilde{\varphi}}^{(k)}\\ 0\end{bmatrix}$ 最远的点近似代替在整个训练样本集 $\boldsymbol{S}$ 所对应的高维特征空间中寻找离球心 $\begin{bmatrix}\boldsymbol{c}_{\tilde{\varphi}}^{(k)}\\ 0\end{bmatrix}$ 最远的点. 文献[27]指出, 当 $|\boldsymbol{S}_{\mathrm{sub}}| = 59$ 时, 最远点以95%的概率落入 $\boldsymbol{S}_{\mathrm{sub}}$, 且时间复杂度为 $O(v^2\cdot|\boldsymbol{X}_k|^2 + 59v\cdot|\boldsymbol{X}_k|)$.

1.4　MvCVM算法性质

MvCVM算法是将两类核化L2-SVM算法应用于多视角学习而得到的一种特殊多视角CCMEB学习算法, 因此, 关于CCMEB核心集的结论仍然适用于MvCVM. 根据文献[17-18, 23-24, 27], 可以得到MvCVM算法的重要性质.

性质 1 对于给定的CCMEB逼近精度阈值ε, MvCVM算法核心集基数和迭代次数上界均为$O(1/\varepsilon)$.

性质 2 对于给定的CCMEB逼近精度阈值ε, MvCVM算法时间复杂度上界为$O(l/\varepsilon^2+1/\varepsilon^4)$, 即MvCVM算法时间复杂度与样本大小$l$呈线性关系.

性质 3 对于给定的CCMEB逼近精度阈值ε, MvCVM算法空间复杂度上界为$O(1/\varepsilon^4)$, 即可以使用存储核心集代替存储所有样本.

正是上述性质使得本文提出的MvCVM算法不仅具备多视角学习的良好特质, 而且还具备对大样本多视角数据集处理的优越性能.

2 实验研究

2.1 实验数据描述

本节通过2个人造多视角数据集和2个真实多视角数据集进行实验来对本文提出的Multi-view L2-SVM和MvCVM算法性能进行评估, 所使用多视角数据集描述如表1和表2所示. 其中: 表1为小样本多视角数据集, 表2为大样本多视角数据集.

表 1 小样本多视角数据集描述

Datasets	View		Sample count	
	Name	#Dimension	#Positive	#Negative
Two moons and Two lines	Two moons	2	150	150
	Two lines	2		
WebKB	Page	3 000	230	821
	Link	1 840		

表 2 大样本多视角数据集描述

Datasets	View		Sample count	
	Name	#Dimension	#Positive	#Negative
Two moons and Two lines	Two moons	2	150 000	150 000
	Two lines	2		
USPS01	View1	225	187 913	153 549
	View2	225		
	View3	225		

人造多视角数据集各个视角参照文献[28]方法随机生成, WebKB数据集参照文献[29]处理, USPS01数据集可由http://www.c2i.ntu.edu.sg/ivor/cvm.html下载得到.

2.2 实验设置

为了说明本文方法对多视角数据分类的良好性能, 通过两组实验进行比较. 第1组实验在表1所示的小样本人造多视角数据集和真实多视角数据集上进行; 第2组实验在表2所示的大样本人造多视角数据集和真实多视角数据集上进行. 通过测试人造多视角数据集来说明本文方法在寻找决策超平面过程中多个视角之间的相互协作而使最终的分类效果趋于一致; 通过测试真实多视角数据集来表明本文方法是一种有效的分类方法. 为了进一步说明本文MvCVM方法对大样本多视角数据集分类表现出良好的性能, 分别在人造大样本多视角数据集和USPS01真实大样本多视角数据集上进行实验.

在实验中, 将不同视角所拼接形成的视角称为混合视角(Hybrid View). 另外, 为了实验的公平性和合理性, 对L2-SVM、Multi-view L2-SVM、SVM-2K[7]、MvCVM和GCVM方法都按照文献[30]方法使用高斯核$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma)$并使用网格搜索对核宽$\sigma$和惩罚因子进行寻优. 其中每个视角核宽参数$\sigma$以训练样本所对应视角的平均$l_2$范数平方$s$为基准, 并在网格$\{s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s, 16s, 32s, 64s\}$中搜索选取. 各视角惩罚因子在网格$\{1, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$中搜索选取. 由定理1的证明可知, 在获取各视角核宽参数和惩罚因子后, 即可自动获取参数η的初始值. 本文的所有实验环境均为Intel(R) Core(TM) i5-4200 M 2.5 GHz, 4 GB内存, Win7 64位专业版, Matlab R2010a 64位.

由于表1和表2多视角数据集含有不平衡数据集, 本文采用几何均值(Gm)[31]准确率对L2-SVM、Multi-view L2-SVM、SVM-2K、MvCVM和GCVM方法的分类结果作出合理的评价.

由式(6)可以看出, 在使用Gm准确率对Multi-view L2-SVM和MvCVM两方法的分类结果进行评价时, 不仅可以对整体(整个多视角数据集或混合视角)作出一致性评价, 而且还可以对构成多视角数据集的各个视角进行评价. 因此, 在下面的具体实验部分, 除了给出Multi-view L2-SVM、SVM-2K和MvCVM三方法在混合视角上的Gm准确率外, 还给出了三方法在各个视角上的准确率.

2.3 小样本多视角数据集实验

由于L2-SVM本质上是单视角分类算法, 在本节小样本多视角数据集实验中, 将L2-SVM算法分别运行于组成多视角数据集的各个视角及混合视角, 而将Multi-view L2-SVM和SVM-2K算法运行于混合视角.

2.3.1 人造小样本多视角数据集实验

图1为L2-SVM、Multi-view L2-SVM和SVM-2K三种算法对表1所示人造双月和双线多视角数据集各运行一次的决策超平面结果.

从图1所示决策超平面可以看出: L2-SVM方法仅将各个视角简单拼接, 不能在构成混合视角的各个视角上获得一致分类结果(见图1(a)和图1(b)); 而本文所提出的Multi-view L2-SVM充分利用了各个视角

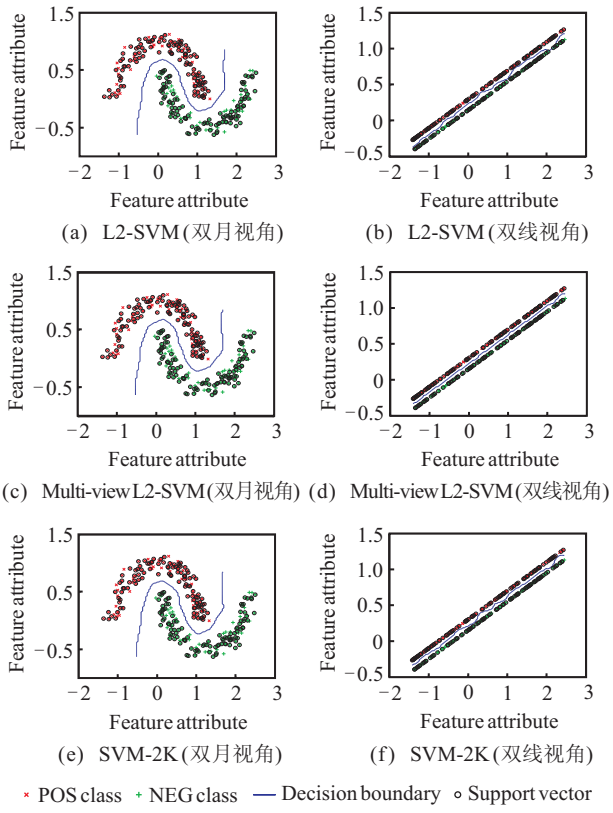

图 1　人造小样本多视角数据集决策超平面结果

之间的差异性与关联性, 使得各个视角得到了趋于一致的分类结果 (见图 1(c) 和图 1(d)), 并且 Multi-view L2-SVM 方法在各个视角上的支持向量数可以不同.

2.3.2　小样本真实多视角数据集实验

在本节实验中, 随机将 70% 的样本作为训练集, 而将剩余的 30% 作为测试集, 实验记录了 3 种算法 10 次运行结果. 表 3 给出了 L2-SVM、Multi-view L2-SVM 和 SVM-2K 三算法对多视角 WebKB 数据集 10 次运行结果的 Gm 准确率平均值及方差, 其中 Gm 准确率方差置于括号中.

表 3　三算法分类准确率比较　%

Method	Page view Gm	Link view Gm	Hybrid view Gm
L2 − SVM	92.564 4	93.422 0	**95.329 2**
	(2.886 9)	(2.493 6)	**(1.371 0)**
Multi-view L2-SVM	95.844 2	91.081 3	**96.674 8**
	(1.065 7)	(2.451 8)	**(0.975 4)**
SVM-2K	93.380 6	93.067 9	**95.966 1**
	(2.112 1)	(1.460 7)	**(1.243 6)**

从表 3 容易看出: 三算法都能在混合视角上获得较高的 Gm 准确率, Multi-view L2-SVM 方法能充分利用各个视角之间的差异性与关联性, 使得最终的学习效果 (分类结果) 在各个视角上趋于一致.

2.4　大样本多视角数据集实验

SVM-2K 是一种小样本多视角算法[6], 而本文旨在提出一种适合于大样本多视角的有效算法, 因此, 本节将通过表 2 所示的大样本多视角数据集来对 Multi-view L2-SVM、MvCVM 和 GCVM 进行比较, 以说明 Multi-view L2-SVM 是一种小样本多视角算法, MvCVM 是一种适合于大样本多视角的有效算法. 在实验中, 随机从大样本多视角数据集中采样大小不等的样本作为训练集, 并随机采样 500 个样本作为测试集, 实验记录了三算法的 Gm 准确率平均值和以秒 (s) 为时间单位的训练时间 Tr-time 平均值, 并用 "—" 表示由于 Matlab 内存溢出错误而导致相应算法不能顺利运行.

2.4.1　人造大样本多视角数据集实验

Multi-view L2-SVM、MvCVM 和 GCVM 三算法在随机采样样本集大小为 6 000 时混合视角各运行一次的决策超平面结果显示于图 2, 表 4 给出了三算法对表 2 所示人造大样本多视角数据集在不同样本集大小时的 10 次运行结果. 本节关于 MvCVM 和 GCVM 的实验是在 CCMEB 逼近精度阈值 ε 为 10^{-5} 时完成的.

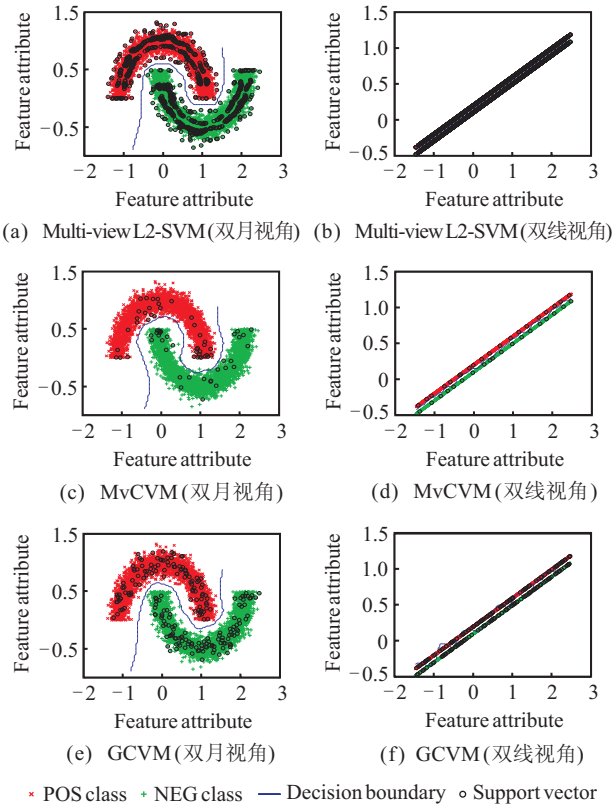

图 2　人造大样本多视角数据集决策超平面结果

与小样本人造多视角数据集实验结果类似, 表 4 记录的结果和图 2 所示的决策超平面同样表明, Multi-view L2-SVM 和 MvCVM 两多视角算法都能充分利用各个视角的差异性与视角之间的关联性, 使得最终的决策超平面或分类结果对各个视角都得到趋于一致的结果, 但由于 MvCVM 和 GCVM 算法是在核心集上求解支持向量, 从而使用 MvCVM 和 GCVM 算

表 4 三算法在人造大样本多视角数据集上实验结果

#Sampled sizes		Multi-view L2-SVM		MvCVM		GCVM	
		Gm/%	Tr-time/s	Gm/%	Tr-time/s	Gm/%	Tr-time/s
2 000	Two moons view	99.960 0		99.544 6		**100.000 0**	0.553 8
		(0.085 0)		(0.579 9)		**(0.000 0)**	(0.174 8)
	Two lines view	**100.000 0**	285.940 0	99.898 9	119.435 9	**100.000 0**	0.316 7
		(0.000 0)	(23.431 0)	(0.254 8)	(86.186 5)	**(0.000 0)**	(0.219 4)
	Hybrid view	**100.000 0**		**100.000 0**		99.387 5	28.836 8
		(0.000 0)		**(0.000 0)**		(0.250 1)	(5.395 8)
6 000	Two moons view	**100.000 0**		99.657 6		**100.000 0**	0.505 4
		(0.000 0)		(0.259 8)		**(0.000 0)**	(0.213 2)
	Two lines view	**100.000 0**	5 563.491 7	99.720 9	90.198 2	**100.000 0**	0.432 1
		(0.000 0)	(269.323 1)	(0.610 3)	(35.724 9)	**(0.000 0)**	(0.400 2)
	Hybrid view	**100.000 0**		**99.961 6**		99.495 6	43.287 2
		(0.000 0)		**(0.121 4)**		(0.320 3)	(13.229 4)
10 000	Two moons view	—		98.720 7		**100.000 0**	0.644 3
				(1.478 6)		**(0.000 0)**	(0.252 2)
	Two lines view	—	—	**100.000 0**	92.226 2	**100.000 0**	0.262 1
				(0.000 0)	(71.710 2)	**(0.000 0)**	(0.200 7)
	Hybrid view	—		**100.000 0**		99.442 3	39.047 1
				(0.000 0)		(0.333 9)	(7.958 8)
20 000	Two moons view	—		99.781 6		99.980 8	0.728 5
				(0.280 6)		(0.060 6)	(0.170 1)
	Two lines view	—	—	99.873 7	130.529 2	**100.000 0**	0.396 2
				(0.399 5)	(99.670 8)	**(0.000 0)**	(0.297 8)
	Hybrid view	—		**99.979 1**		99.365 2	52.645 7
				(0.066 2)		(0.305 9)	(12.174 3)
100 000	Two moons view	—		99.669 2		**100.000 0**	1.335 4
				(0.4824)		**(0.000 0)**	(0.415 1)
	Two lines view	—	—	99.903 8	113.890 1	**100.000 0**	1.090 4
				(0.3044)	(49.849 9)	**(0.000 0)**	(0.658 8)
	Hybrid view	—		**99.942 4**		99.474 0	54.720 5
				(0.182 3)		(0.207 1)	(13.183 3)
300 000	Two moons view	—		99.672 1		**100.000 0**	3.067 0
				(0.611 8)		**(0.000 0)**	(0.309 4)
	Two lines view	—	—	99.903 8	281.709 7	**100.000 0**	2.681 7
				(0.304 4)	(372.919 6)	**(0.000 0)**	(1.154 3)
	Hybrid view	—		**100.000 0**		99.526 2	61.628 2
				(0.000 0)		(0.247 7)	(15.100 1)

法进行求解所得支持向量数较 Multi-view L2-SVM 算法求解少. 由表 4 可见, 当用于训练的样本数大于 6 000 时, Multi-view L2-SVM 算法在本文实验环境中已无法顺利运行.

2.4.2 真实大样本多视角数据集实验

本节将通过如表 2 所示的 USPS01 真实大样本多视角数据集来说明本文提出的 MvCVM 算法对大样本多视角数据进行分类的优越性和有效性. 本文对多视角 USPS01 数据集视角构造方法是: 将数据集 675 个特征依次平均分成 3 组特征子集, 每组特征子集构成数据集的一个视角, 每组特征子集基数均为 225. 在本节实验中, CCMEB 逼近精度阈值 ε 取为 10^{-4}.

表 5 给出了三算法在大样本多视角 USPS01 数据集上分别运行 5 次的实验结果. 从表 5 记录的结果可以看出: Multi-view L2-SVM 算法和 MvCVM 算法都能充分利用各个视角的差异性与视角之间的关联性, 在混合视角上获得了较高的 Gm 准确率 (一致性); MvCVM 算法既是一种适于小样本的多视角学习算法, 又是一种适于大样本的多视角学习算法.

3 结 论

Multi-view L2-SVM 算法是将两类核化 L2-SVM 算法应用于多视角学习而提出的多视角支持向量机算法, 方法独特. 由于 Multi-view L2-SVM 分类器等价于一 CCMEB 问题, 可将 GCVM 算法用于 Multi-view L2-SVM 分类器以解决大样本多视角分类问题. 基于此, 本文提出了适合于大样本的多视角学习算法 MvCVM. 在小样本和大样本数据集上的实验表明, Multi-view L2-SVM 算法和 MvCVM 算法都能有效利用各视角的差异性和关联性来改善分类器学习性能, 使得各视角分类结果趋于一致. 在 MvCVM 算法实现中, 本文直接采用概率加速近似算法, 而没有进行深入的理论分析; 同时, MvCVM 算法中 CCMEB 逼近精

表 5　三算法在多视角 USPS01 数据集上实验结果

#Sampled sizes		Multi-view L2-SVM		MvCVM		GCVM	
		Gm/%	Tr-time/s	Gm/%	Tr-time/s	Gm/%	Tr-time/s
1 000	View 1	85.073 3		70.398 9		61.651 2	24.457 8
		(0.995 0)		(3.886 6)		(1.494 8)	(7.271 6)
	View 2	95.528 9		94.568 2		**96.090 9**	102.904 5
		(3.972 6)	192.864 0	(1.297 1)	319.933 1	**(1.019 0)**	(25.763 6)
	View 3	86.337 3	(12.249 5)	78.175 5	(40.844 3)	75.406 4	28.245 5
		(1.899 0)		(2.382 4)		(1.753 7)	(2.590 7)
	Hybrid view	**97.770 4**		**95.003 7**		90.278 5	86.015 8
		(0.478 2)		**(0.478 4)**		(0.773 6)	(10.689 0)
3 000	View 1	86.123 2		71.247 8		62.035 1	54.715 8
		(2.086 0)		(3.781 6)		(1.168 7)	(8.936 1)
	View 2	96.461 1		94.427 7		**94.999 3**	197.547 2
		(1.720 4)	3 773.018 3	(1.387 6)	541.963 1	**(1.799 3)**	(51.598 8)
	View 3	88.102 7	(115.197 7)	77.147 0	(83.747 2)	75.612 2	61.455 0
		(2.424 6)		(2.939 7)		(2.260 4)	(18.462 7)
	Hybrid view	**98.182 5**		**94.790 5**		91.265 3	189.394 6
		(0.261 9)		**(0.746 2)**		(1.615 0)	(35.472 6)
4 000	View 1	—		68.453 9		64.819 3	64.082 1
				(2.208 9)		(2.379 7)	(19.622 3)
	View 2	—		94.263 2		**96.415 9**	297.163 2
			—	(1.027 5)	742.596 0	**(1.585 6)**	(124.103 5)
	View 3	—		77.001 0	(149.536 3)	74.990 5	98.511 5
				(1.927 2)		(2.090 3)	(23.794 2)
	Hybrid view	—		**94.714 9**		91.887 0	259.033 4
				(0.611 3)		(1.381 1)	(42.416 5)
10 000	View 1	—		72.888 7		67.198 5	181.320 0
				(3.714 0)		(3.031 7)	(104.378 8)
	View 2	—		93.917 8		**93.158 4**	1126.052 7
			—	(1.098 1)	1 330.832 1	**(2.725 5)**	(311.588 5)
	View 3	—		77.387 1	(569.510 8)	75.226 3	325.165 4
				(3.317 3)		(2.055 3)	(110.304 8)
	Hybrid view	—		**95.077 9**		90.882 4	638.172 0
				(1.013 7)		(1.297 1)	(111.174 4)
100 000	View 1	—		71.128 9		62.548 3	2657.455 2
				(1.889 3)		(3.530 2)	(495.003 9)
	View 2	—		94.934 8		**94.919 0**	18 698.450 0
			—	(0.878 2)	25 748.411 2	**(1.720 8)**	(2 552.487 2)
	View 3	—		78.872 4	(8 335.150 4)	75.544 4	3 623.859 5
				(1.778 5)		(1.001 0)	(1 244.405 9)
	Hybrid view	—		**95.217 2**		90.274 6	8 254.084 7
				(1.100 3)		(1.433 9)	(3 435.369 6)
341 462	View 1	—		69.932 6		63.972 2	8 195.163 1
				(2.138 4)		(4.396 4)	(1 358.704 7)
	View 2	—		93.663 3		**94.974 4**	28 328.789 4
			—	(1.014 5)	92 516.926 6	**(1.886 1)**	(14 133.486 2)
	View 3	—		76.852 9	(24 333.442 4)	74.676 2	20 845.165 6
				(1.380 7)		(0.969 2)	(5 591.800 4)
	Hybrid view	—		**95.234 1**		90.122 3	49 320.321 4
				(0.645 2)		(1.117 2)	(7 046.042 6)

度阈值也会影响算法的效率和测试准确率, 如何选取 CCMEB 逼近精度阈值本文也未进行深入探讨. 上述问题有待于继续深入研究, 并将 MEB 的方法和理论应用于未来的研究中.

参考文献(References)

[1] Sun S L. A survey of multi-view machine learning[J]. Neural Computing and Applications, 2013, 23(7/8): 2031-2038.

[2] Li G, Chang K, Hoi S C H. Multi-view semi-supervised learning with consensus[J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(11): 2040-2051.

[3] Sun S L. Multi-view Laplacian support vector machines[C]. Proc of the 7th Int Conf on ADMA. Berlin: Springer, 2011: 209-222.

[4] Zhang Q, Sun S. Multiple-view multiple-learner active

learning[J]. Pattern Recognition, 2010, 43(9): 3113-3119.

[5] Li G, Hoi S C H, Chang K. Two-view transductive support vector machines[C]. Proc of the SIAM Int Conf on Data Mining. Columbus, 2010: 235-244.

[6] Sun S L, Shawe-Taylor J. Sparse semi-supervised learning using conjugate functions[J]. J of Machine Learning Research, 2010, 11(9): 2423-2455.

[7] Farquhar J, Hardoon D, Meng H, et al. Two view learning: SVM-2K, theory and practice[C]. Proc of Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2005: 355-362.

[8] Sindhwani V, Niyogi P, Belkin M. A co-regularization approach to semi-supervised learning with multiple views[C]. Proc of the ICML 2005 Workshop on Learning With Multiple Views. Bonn, 2005: 74-79.

[9] Ando R K, Zhang T. Two-view feature generation model for semi-supervised learning[C]. Proc of the 24th Int Conf on Machine Learning. Corvallis, 2007: 25-32.

[10] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]. Proc of the Eleventh Annual Conf on Computational Learning Theory. Madison, 1998: 92-100.

[11] Collins M, Singer Y. Unsupervised models for named entity classification[C]. Proc of the Joint SIGDAT Conf on Empirical Methods in Natural Language Processing and Very Large Corpora. Maryland, 1999: 100-110.

[12] Muslea I, Minton S, Knoblock C A. Selective sampling with redundant views[C]. Proc of AAAI-2000. Austin: AAAI Press, 2000: 621-626.

[13] Chaudhuri K, Kakade S M, Livescu K, et al. Multi-view clustering via canonical correlation analysis[C]. Proc of the 26th Annual Int Conf on Machine Learning. Montreal, 2009: 129-136.

[14] De S V R. Spectral clustering with two views[C]. Proc of ICML Workshop on Learning with Multiple Views. Bonn, 2005: 20-27.

[15] Kailing K, Kriegel H P, Pryakhin A, et al. Clustering multi-represented objects with noise[C]. Proc of PAKDD. Berlin: Springer, 2004: 394-403.

[16] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[17] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. J of Machine Learning Research, 2005, 6(4): 363-392.

[18] Tsang I W, Kwok J T, Zurada J M. Generalized core vector machines[J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126-1140.

[19] 胡文军, 王士同, 王娟, 等. 一般化最小包含球的大样本快速学习方法[J]. 自动化学报, 2012, 38(11): 1831-1840.
(Hu W J, Wang S T, Wang J, et al. Fast learning of generalized minimum enclosing ball for large datasets[J]. Acta Automatica Sinica, 2012, 38(11): 1831-1840.)

[20] 钱鹏江, 王士同, 邓赵红. 大数据集快速均值漂移谱聚类算法[J]. 控制与决策, 2010, 25(9): 1307-1312.
(Qian P J, Wang S T, Deng Z H. Fast mean shift spectral clustering on large data sets[J]. Control and Decision, 2010, 25(9): 1307-1312.)

[21] 胡文军, 王士同, 邓赵红. 适合大样本快速训练的最大夹角间隔核心集向量机[J]. 电子学报, 2011, 39(5): 1178-1184.
(Hu W J, Wang S T, Deng Z H. Maximum vector-angular margin core vector machine suitable for fast training for large datasets[J]. Acta Electronica Sinica, 2011, 39(5): 1178-1184.)

[22] Chung F L, Deng Z H, Wang S T. From minimum enclosing ball to fast fuzzy inference system training on large datasets[J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 173-184.

[23] Deng Z H, Chung F L, Wang S T. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation[J]. Pattern Recognition, 2008, 41(4): 1363-1372.

[24] Bădoiu M, Clarkson K L. Optimal core sets for balls[J]. Computational Geometry, 2002, 40(1): 14-22.

[25] Schölkopf B, Platt J C, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[26] Kumar P, Mitchell J S B, Yildirim E A. Approximate minimum enclosing balls in high dimensions using core-sets[J]. J of Experimental Algorithmics, 2003, 8(1): 1-29.

[27] Smola A J, Schölkopf B. Sparse greedy matrix approximation for machine learning[C]. Proc of the 17th Int Conf on Machine Learning. California, 2000: 911-918.

[28] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. J of Machine Learning Research, 2006, 7(11): 2399-2434.

[29] Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: From transductive to semi-supervised learning[C]. Proc of the 22nd Int Conf on Machine Learning. Bonn, 2005: 824-831.

[30] Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.

[31] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection[C]. Proc of the 14th Int Conf on Machine Learning. Nashville, 1997: 179-186.

(责任编辑：李君玲)