Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2013, 2, 67-70
http://dx.doi.org/10.12677/airr.2013.21012 Published Online February 2013 (http://www.hanspub.org/journal/airr.html)

New Binary Classifier P2M-SVM*

Jinjin Liang[1#], De Wu[2]

[1]Department of Mathematical Sciences, Xi'an Shiyou Univerisity, Xi'an
[2]School of Computer Sciences, Xidian University, Xi'an
Email: #myonlyonly@126.com

Received: Jan. 18th, 2013; revised: Jan. 31st, 2013; accepted: Feb. 6th, 2013

Abstract: A semi-supervised binary support vector machine (SVM) is proposed based on possibilistic two-means (P2M) clustering. First, divide the unlabeled data using PCM; then, train the labeled data using SVM. Experiments on artificial and UCI data show the superiority over existing algorithm. P2M-SVM utilizes both the robustness of P2M for binary clustering and the strong generalization ability of SVM for classification thus increases the classification accuracy of traditional clustering and reduces the cost of sample collecting of the SVM.

Keywords: P2M; Semi-Supervised; SVM; Robustness; Generalization Ability

新型二分类支持向量机 P2M-SVM*

梁锦锦[1#]，吴　德[2]

[1]西安石油大学理学院，西安
[2]西安电子科技大学计算机学院，西安
Email: #myonlyonly@126.com

收稿日期：2013 年 1 月 18 日；修回日期：2013 年 1 月 31 日；录用日期：2013 年 2 月 6 日

摘　要：提出基于可能性二均值聚类(Possibilistic Two Means, P2M)的二分类支持向量机(Support Vector Machine, SVM)。该算法先用 P2M 对未知类别的二分类数据进行划分，然后利用支持向量机对划分后的数据进行训练。人造数据和 UCI 数据上的分类实验表明，该算法综合利用了 P2M 聚类的稳健性和 SVM 分类的强泛化能力，提高了传统聚类的分类精度并降低了 SVM 的类别采集代价。

关键词：可能性二均值聚类；半监督二分类支持向量机；全局最优；稳健性；泛化能力

1. 引言

基于结构风险最小化原则建立起来的小样本机器学习方法）——支持向量机(Support Vector Machine, SVM)由 Vapnik 等提出，是一种基于统计学习理论的有监督学习方法[1,2]。由于支持向量机具有拟合精度高，选择参数少，推广能力强和全局最优等特点，能够较好的解决小样本，高维数，非线形，局部极小等问题，而成为机器学习领域新的研究热点，并被用于人脸识别，文本分类，手写体识别和蛋白质结构预测等领域[3,4]。由于大量的信息以数据的形式产生并被无标志的保存，且对很多现实问题数据进行手工分类是不可行的，这就限制了标准 SVM 的应用。

聚类作为一种无监督分类方法，按照一定的规则将数据分成不同的簇，使得同一簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。传统的 C 均值硬划聚类将每个待处理对象严格的划分到每个类中，隶属度不是 1 就是 0，从而不能真正反映对象和类的实际关系[5,6]。Zedeh 提出的模糊集理论为软化分提供了有利的分析工具，其中以 Dunn 提出并由

*基金项目：国家自然科学基金(No.60974082)，陕西省教育厅专项研究基金(2010JK773)，西安石油大学专项科研基金(Z10027)。
#通讯作者。

Bezdek 加以推广的模糊 C-均值(FCM)得到了广泛而且较成功的应用[7,8]。由于 FCM 要求隶属度归一，其缺陷是容易对数据中的噪声和孤立点赋予较大的隶属度而得不到好的聚类效果。Krishnapuram 等通过放宽 FCM 的概率约束，提出 PCM 算法，使得隶属度真正代表样本隶属于某一类的可能性，进而提高对噪声和野值的鲁棒性[9,10]。

如何综合利用 SVM 的泛化能力和聚类的无监督学习能力构造具有良好学习能力的分类器是本文工作的出发点。以二分类问题为例，本文先采用 P2M 获得无标志样本的类别指标，证明了解的全局最优性，再对新划分数据进行支持向量机训练。人工数据和 UCI 数据上的模拟实验表明了该算法的有效性和优越性。全文组织结构如下：第一部分阐述 PCM 聚类描述的基础，并导出分类器 P2M-SVM；第二部分给出数值试验来验证新算法的可行性和有效性。第三部分是结论，给出进一步要做的工作。

2. 二分类支持向量机 P2M-SVM

2.1. PCM 聚类描述

设数据集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$，$\boldsymbol{U}=\{u_{ik}\}_{c\times n}$ 是一个隶属度矩阵，$v=\{v_1,v_2,\cdots,v_c\}$ 是 $c\left(v_i\in R^s,2\le c\le n\right)$ 个聚类中心。

C-均值算法把 n 个向量 $x_i\left(i=1,2,\cdots,n\right)$ 分成 C 个簇 $G_i\left(i=1,2,\cdots,c\right)$，并求得每个簇的聚类中心，使得簇内方差和 $J(u,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\left\|x_k-v_i\right\|^2$ 达到最小 $\left(\sum_{i=1}^{c}u_{ik}=1,u_{ik}\in\{0,1\}\right)$。由于其隶属度非 0 即 1，不能反映数据点和类中心的实际关系。Bezdek 通过在目标函数中增加模糊权重指数 m，并修改隶属度函数 $u_{ik}\in(0,1)$ 得到模糊 C-均值 FCM。由于 FCM 对噪声敏感，Krishnapuram 等放宽其概率约束，提出 PCM 算法使得隶属度真正代表样本隶属于某一类的可能性，以提高对噪声和野值的鲁棒性。记学习因子为 $\eta_i\ge 0$，$d_{ik}=\left\|x_k-v_i\right\|$，PCM 的优化目标为：

$$\min\ J(u,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^{m}d_{ik}^{2}+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}\left(1-u_{ik}\right)^{m}$$

$$s.t.\quad 0\le\sum_{i=1}^{c}u_{ik}\le C,\qquad u_{ik}\in(0,1)$$

根据 Lagrange 法对约束引入拉格朗日乘子，对优化目标的拉格朗日函数关于隶属度 u_{ik} 和类中心 v_i 求偏导，令偏导数为零可得相应的更新规则。PCM 具体步骤如下：

1) 初始化聚类中心 V^0；设定学习因子 $\eta_i\ge 0(i=1,2)$，聚类个数 C，权重指数 $m\left(2\le m<\infty\right)$，阈值 ε 和最大迭代次数 $T>0$；置迭代计数器 $l=0$。

2) 更新 $u_{ik}^{(l+1)}$。$u_{ik}^{(l+1)}=\left(1+\left(d_{ik}^2/\eta_i\right)^{1/m-1}\right)^{-1}$。

3) 更新 $v_i^{(l+1)}$。$v_i^{(l+1)}=\sum_{k=1}^{n}u_{ik}^{(l+1)}x_k\Big/\sum_{k=1}^{n}u_{ik}^{(l+1)}$。

如果 $\max_i\left\|v_i^{(l+1)}-v_i^{(l)}\right\|<\varepsilon$ 或者 $l>T$，则停止；否则 $l=l+1$，转至步骤 2)。

2.2. 分类器 P2M-SVM

随网络的发展，大量的信息以数字数据的形式产生并被无标志保存，且对很多现实问题进行手工分类不可能。传统的二分类 SVM 在识别前需要利用已知训练集进行训练，这无疑限制了自身的应用范围。P2M-SVM 首先在训练集上执行 P2M 聚类，收敛后得到所有样本的类别指标及相应于目标类的隶属度，然后调用标准 SVM 得到分类决策。传统的 PCM 对隶属度和簇类中心进行交替优化，如下定理 1 保证了解的局部收敛性[10]。

定理 1 PCM($C\ge 2$)是局部最优算法[10]。

该算法的一个重要步骤是运行 P2M 前初始聚类中心的选取和参数的设置：初始聚类中心决定了算法的收敛速度；权重指数 m 决定了数据集中所有点属于某簇的概率，随 m 的增加其归属该簇的概率也增加；学习因子 η_i 决定了临界样本(具有最大模糊性)与簇中心的距离，随 η_i 的增加其归属该簇的隶属度 u_{ik} 也增加。

本文以标准 F2M 算法产生的聚类中心作为初始聚类中心，通过多次实验设定权重指数 $m=2$，设置学习因子 η_i 正比于簇内平均模糊距离(见以下定义)。分别记 F2M 的聚类中心和隶属度为 $v_i\left(i=1,2\right)$ 和 $u_{ik}\left(i=1,2;k=1,\cdots,n\right)$。记 $d_{ik}=\left\|x_k-v_i\right\|(i=1,2;k=1,\cdots,n)$，则

$$\eta_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}d_{ik}^{2}}{\sum_{k=1}^{n}u_{ik}^{m}}$$

算法收敛后再次执行 P2M 算法得到最终聚类结果。

3. 数值实验

为验证 P2M-SVM 性能，选取正态分布数据和 UCI 数据进行实验。实验均在 CPU 为 P4，3.06 GHz，内存为 0.99 GB 的 PC 机上进行，采取 Matlab 7.01 实现。以去掉类别标志的样本作为训练集，P2M-SVM 先通过 P2M 聚类得到样本的类别指标，然后采用 SVM 训练得到最终的决策规则，对比已有算法的数值实验如下。

例 1 仿真试验

首先表明好的初始划分的重要性。产生两类完全可分的正态分布点，并对其中某类加入部分噪声。设定模糊权重 $m=2$，依次列出 CM，FCM 和 PCM 的聚类结果如下($C=2$)。

显然 P2M 具有最优的划分，更接近客观分布。预计基于 P2M 的支持向量机具有良好的表现见后续实验(图 1)。

例 2 UCI 上两分类数据集 Breast Cancer 和 Diabetis。

去掉类别标记，Breast Cancer 为含有 277 个样本的 9 维数据，Diabetis 为含有 768 个的样本的 8 维数据。选取前者的 200 个样本作为训练集，其余 77 个作为测试集；选取后者的 400 个作为训练集，其余 368 个作为测试集。首先采用 P2M 获得无标志样本的类别，然后进行 SVM 训练。选定径向基核函数，对比 CM，FCM，P2M 和 P2M-SVM 的结果。

显然 P2M-SVM 提高了已有聚类算法的训练精度和测试精度(表 1)！由于传统聚类算法根据测试样本到聚类中心的最小距离判别样本的归属，而 P2M 采用 SVM 求取已知类别样本的最优分类超平面并根据测试样本到最优超平面的距离判别其归属；后者较强的泛化能力导致较高的训练和测试精度！

由上表还可以观察到，作为一种特殊的支持向量机，P2M-SVM 对核参数变化较为敏感。为进一步验证 P2M-SVM 的性能，本文细化径向基核参数的取值，以 Breast Cancer 为例对比相同惩罚因子下分类良好的 SVM 与本文算法的分类精度随径向基核参数变化趋势见图 2。

从图 2 可以看出：1) 随径向基核参数的增大，P2M-SVM 和 SVM 的分类精度也随之上升；当核参数增大到一定数值，两者的分类精度随核参数变化影响不大；2) 对于选择适当的核函数(如核参数介于 0.01

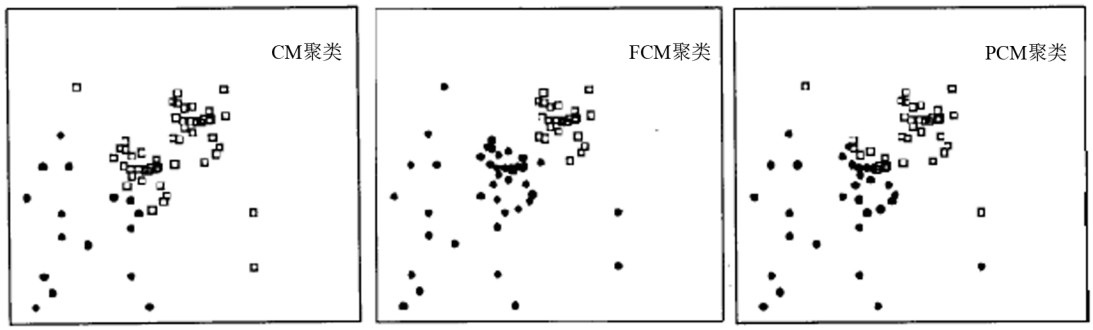

Figure 1. Curve: Clustering results of various algorithms
图 1. 不同算法的聚类结果

Table 1. Performances comparisons of various algorithms
表 1. 不同算法的结果比较

	算法	CM	FCM	P2M	P2M-SVM		
					$\sigma=0.01$	$\sigma=0.1$	$\sigma=0.5$
训练精度	Breast Cancer	46.50%	56.00%	63.81%	75.33%	80.56%	78.14%
	Diabetis	66.45%	67.74%	67.96%	76.32%	80.65%	79.13%
测试精度	Breast Cancer	23.38%	49.35%	55.63%	25.71%	63.07%	60.41%
	Diabetis	31%	46%	50.17%	13.69%	65.14%	65.63%

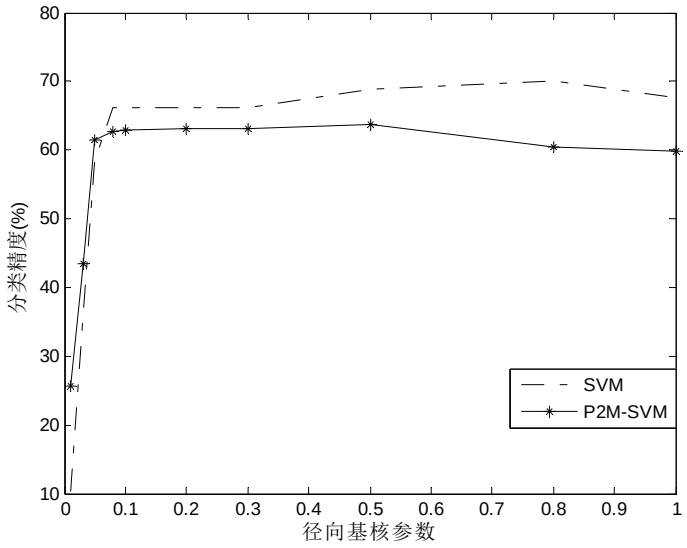

Figure 2. Curve: Variations of accuracies with radial basis kernel parameter
图 2. 不同算法的分类精度随径向基核参数的变化

和 0.3 之间)，两者的分类精度相差不大；而对核参数的其他取值，两者分类精度的差也小于 10%；从而 P2M-SVM 为半监督分类提供了新的方法！

4. 结论

本文提出了一种新型二分类支持向量机 P2M-SVM，并证明了解的全局最优性。由于综合利用了 P2M 聚类的稳健性和 SVM 分类的强泛化能力，该算法有效提高了传统聚类方法 CM，FCM，PCM 的分类精度，并且降低了有监督 SVM 的样本类别采集代价。UCI 数据集的二分类实验表明：对合适选择的核参数，P2M-SVM 的分类精度与 SVM 精度相当；对于其他的核参数，两者分类精度的差小于 10%，从而为无标志样本的二分类问题提供了一种新的方法。值得注意的是，虽然 P2M-SVM 的训练精度低于 SVM，但是两者的分类精度却相差不大。其原因可能是由于 SVM 出错点多分布在分类边界附近，P2M-SVM 得到了对中心样本的正确划分，从而保证了用于 SVM 分类的泛化能力！下一步的工作是将该算法用于多分类问题并根据 PCM 的聚类结果压缩训练集，在不影响聚类结果的基础上选择隶属度较大的样本参与训练！

参考文献 (References)

[1] N. Vanik. 统计学习理论[M]. 许建华, 张学工, 译. 北京: 电子工业出版社, 2004.
[2] N. Cristianini, J. S. Taylor. An introduction to support vector machines. Cambridge University Press, Cambridge, 2000.
[3] V. N. Vapnik. An overview of statistical learning theory. IEEE Transactions on NN, 1999, 10(3): 988-999.
[4] Y. Jin, Y. Ma and L. Zhao. A modified self-training semi-supervised SVM algorithm. Proceedings of the International Conference on Communication Systems and Network Technologies, 2012: 224-228.
[5] K. L. Wu, M. S. Yang. Alternative c-means clustering algorithms. Pattern Recognition, 2002, 35(10): 2267-2278.
[6] 张敏, 于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004, 15(6): 858-868.
[7] J. C. Bezdek. Pattern recognition with fuzzy objective function algorithms. New York: Plenum, 1981.
[8] J. C. Dunn. Some recent investigations of a new fuzzy partion algorithm and its application to pattern classification problems. Journal of Cybernetics, 1974, 4: 1-15.
[9] R. Krishnarpuram, J. Keller. A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98-110.
[10] N. Pal, K. Pal, J. Keller and J. Bezdek. A possinilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems, 2005, 13(4): 517-530.