DOI:10.3969/j.issn.1673-4785.201209062
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130515.0930.006.html

基于加权聚类质心的 SVM 不平衡分类方法

胡小生,钟勇
(佛山科学技术学院 电子与信息工程学院,广东 佛山 528000)

摘　要:不平衡数据分类是机器学习研究的热点问题,传统分类算法假定不同类别具有平衡分布或误分代价相同,难以得到理想的分类结果.提出一种基于加权聚类质心的 SVM 分类方法,在正负类样本上分别进行聚类,对每个聚类,用聚类质心和权重因子代表聚类内样本分布和数量,相等类别数量的质心和权重因子参与 SVM 模型训练.实验结果表明,该方法使模型的训练样本具有较高的代表性,分类性能与其他采样方法相比得到了提升.

关键词:机器学习;不平衡数据分类;聚类质心;支持向量机
中图分类号:TP181　**文献标志码**:A　**文章编号**:1673-4785(2013)03-0261-05

中文引用格式:胡小生,钟勇.基于加权聚类质心的 SVM 不平衡分类方法[J].智能系统学报, 2013, 8(3): 261-265.
英文引用格式:**HU Xiaosheng, ZHONG Yong. Support vector machine imbalanced data classification based on weighted clustering centroid[J]. CAAI Transactions on Intelligent Systems**, 2013, 8(3): 261-265.

Support vector machine imbalanced data classification based on weighted clustering centroid

HU Xiaosheng, ZHONG Yong
(College of Electronic and Information Engineering, Foshan University, Foshan 528000, China)

Abstract: Classification of imbalanced data has become a research hot topic in machine learning. Traditional classification algorithms assume that different classes have balanced distribution or equal misclassification cost, thus, making it hard to get ideal result of classifications. A support vector machine (SVM) classification method based on weighted clustering centroid was proposed in this paper. First, unsupervised clustering was applied to the positive and negative samples respectively to extract the clustering centroid of each clustering, which was represented the most in compactness of the clustering sample. Next, all clustering centroids formed a new set of balance training. In order to minimize the information loss during clustering, each clustering centroid was associated with a weight factor that was defined proportional to the number of samples of the class. Finally, all clustering centroids and weight factors participated in the training of the improved SVM model. Experimental results show that the proposed method can make the sample selected from model train sets more typical and improve the classification performance better than other sampling techniques for dealing with imbalanced data.

Keywords: machine learning; imbalanced data classification; clustering centroid; support vector machine

不平衡分类是目前机器学习和数据挖掘领域的研究热点问题之一,在现实中存在许多实际应用.不平衡数据集是指其中一类(多数类、负类)样本远多于另一类(少数类、正类),由于传统学习算法假定或者期望数据集具有平衡类分布或相等的类误分代价,因此这些算法不能有效表现数据的分布特征,从而导致少数类的分类性能低下.然而在诸如欺诈检测、医疗诊断等实际应用中,少数类是人们关注的类别,往往具有很高的误分代价.

常用的处理不平衡分类问题方法有基于数据层面的方法和基于算法层面的方法[1].基于数据层面利用重采样技术,包括欠采样[2]和过采样[3],使数据集达到平衡.基于算法层面的方法针对的是分类算法,代价敏感学习(cost-sensitive learning)、主动学习、集成学习以及单类别学习等方法,是处理不平衡数据集的常见算法[4-6].这 2 种层面方法各有优缺

收稿日期:2012-09-27.　**网络出版日期**:2013-05-15.
基金项目:佛山市科技发展专项资金资助项目(2011AA100061);佛山市产学研专项资金资助项目(2012HC100272);佛山市教育局智能评价指标体系研究项目(DX20120220).
通信作者:胡小生.E-mail: happyhxs@tom.com.

点,基于数据层面的方法使用范围广泛,而基于算法层面的方法更适合某些特殊领域.支持向量机(support vector machine, SVM)是基于统计学习理论和结构风险最小化原则的机器学习方法,已经成为当前数据挖掘领域最好的方法之一.当样本类别分布均衡时,支持向量机方法能够取得较高的分类精度,然而应用于不平衡数据集时,其分类性能会大大降低.近年来,有许多文献提出了改进的 SVM 算法来处理不平衡分类问题.文献[7]提出在 SVM 中使用不同的惩罚因子处理不平衡分类问题,即加权 SVM(SVM-weight),在实际应用中有明显效果;文献[8]将 veropoulous 的不同惩罚因子方法同 SMOTE 相结合处理不平衡问题;文献[9]提出核边界校准(kernel boundary alignment, KBA)方法,通过调整核函数边界使得 SVM 更适合处理不平衡数据集;文献[10]提出一种基于 SVM 的主动学习方法,该方法选择离当前分类超平面最近的最富信息样例,重新训练 SVM,能够避免搜索全部数据,但是搜索最富信息样例的过程计算量很大.

本文借鉴加权 SVM 思想,提出一种基于加权聚类质心的 SVM 不平衡分类方法(weighted clustering centroid support vector machine, WCC-SVM).该方法首先分别在正负类样本上进行 K 均值聚类,2 类样本聚类数量相同,之后对于每个聚类,得到聚类质心及其关于质心的统计概要信息,此概要信息用权重因子表示,所有的聚类质心组成新训练集,并且聚类质心的权重因子通过与样本惩罚项相结合的方式修改 SVM 目标函数,得到最终的分类学习模型.

1 K 均值聚类及数据预处理

对于不平衡数据集,对多数类样本进行欠采样和对少数类进行过采样均能改变数据分布,使数据达到平衡,但是这 2 种方法都有缺点,过采样容易使分类器学习到的决策域变小,从而可能导致过度拟合问题;欠采样由于删除部分训练样例,会引起信息丢失.为了减少训练样本集中的类样本数量,对类样本进行划分是一个很好的思路.本文选取 K 均值聚类方法,将训练集中的正负类样本分别聚类为 K 个不相交的子集,对于类别分布极端不平衡的数据集,可仅仅在多数类样本上进行 K 均值聚类,其中 K 值与少数类样本数量相等.样本聚类后,与欠采样方法不同,不是在每个聚类子集内选择一部分样本,而是仅仅用聚类质心来代表某个子集,虽然聚类质心最能够表征一个聚类样本的特征,但是由一个聚类质心代表一个聚类内所有样本,同样也不可避免造成信息丢失.为此,还需提供聚类质心的统计概要信息,用权重因子 p 表示,如果类样本数量为 N,某个聚类子集内样本数量为 n_i,则

$$p_i = \frac{n_i}{N}, \sum p_i = 1.$$

由于 K 均值聚类以及 SVM 分类都只能处理数值型的属性,但实际应用中的数据集具有数值型、分类型等属性,因此需要进行数据预处理.对于数值型的属性,为了消除大的数值在 SVM 目标函数中大的影响,需要进行[0,1]标准化,按照式(1)进行转换:

$$a_j = \frac{a_j - a_{\min}}{a_{\max} - a_{\min}}. \tag{1}$$

对于分类型的属性,采样二进制编码方式进行转换,将分类型的属性转换为若干个取值 0 和 1 的属性.例如,对于有姓名、性别 2 个属性的某样本数据“张三,男”,经过二进制编码转换,变为有姓名、性别_男、性别_女 3 个属性,相应的样本数据信息为“张三,1, 0”.获取聚类质心及其权重因子的具体流程如下.

输入:训练数据集 D

1)S = pre_process[D]; //数据预处理

2)train[S] = 选取数据集 S 中的 80%样本,test[S]=S 中剩余的 20%样本;

3)确定聚类的 K 值,计算 train[S]中正类样本数量 N,如果 N<1 000,令 K=N;否则取 K 的值接近 N,例如 K=0.9N 或者 K=0.95N;

4)在正负类样本中分别进行 K 均值聚类算法,分别得到 K 个不相交子集;

5)对正负类样本分别计算 K 个聚类质心 C_i 及权重因子 p_i;

6)输出:K 个正类聚类质心 C_i 和权重因子 p_i,K 个负类聚类质心 C_i 和权重因子 p_i.

在步骤 3)中,根据数据集 train[S]中正类样本数量 N 来决定聚类的 K 值,当 N 值较小时,为了不至于使接下来参与支持向量机训练的正类样本空间变小,取 K=N,即对正类样本不进行聚类压缩,此时每一个正类样本即为聚类质心,其权重因子 p=1/K.而如果当正类样本数量较多时,样本空间有足够的代表性,适当地对正类样本进行聚类压缩,在分类性能差异不明显的情况下,能够减少正、负类样本聚类时间以及支持向量机模型的训练时间.

2 加权聚类质心的 SVM 分类

经过 K 均值聚类得到类别数量相等的聚类质心及其权重因子,这 2K 个聚类质心及其对应的权

重因子一起组成新的训练集,参与加权 SVM 模型训练,现简要介绍其原理.

考虑二分类问题,给定训练样本集 D:

$$D = \{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i \in \mathbf{R}^N, y_i \in \{1, -1\}\},$$
$$i = 1,2,\ldots,K.$$

支持向量机学习器的目标是求解最优分类超平面,也即最大化 2 类样本之间的距离.不失一般性,考虑训练样本为非线性,通过引入非线性变换 $\boldsymbol{\phi}: \mathbf{R}^N \rightarrow H$ 将输入空间 $\mathbf{R}^N$ 映射到高维特征空间 H,从而将非线性问题转变为线性问题.当线性可分时,求解最优分类超平面可归纳为式(2)所示的二次规划问题:

$$\min\ Q(\boldsymbol{w}, b) = \frac{1}{2} \| \boldsymbol{w} \|^2,$$
$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) - 1 > 0,\ i = 1,2,\ldots,K. \tag{2}$$

式中:$\boldsymbol{w}$ 为权重向量,b 为阈值.当线性不可分时,引入惩罚因子 C,对错分样本进行惩罚,此时式(2)重写为:

$$\min\ Q(\boldsymbol{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{K} \xi_i,$$
$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) > 1 - \xi_i,$$
$$\xi_i > 0, i = 1,2,\ldots,K.$$

式中:ξ_i 为松弛变量,用来度量样本分类错误代价.

文中将聚类质心对应的权重因子 p 引入目标函数中,则 $p_i\xi_i$ 表示给样本分类错误代价设置一个权重.因此,分类超平面目标函数可修改为:

$$\min\ Q(\boldsymbol{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{K} p_i\xi_i,$$
$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) > 1 - \xi_i,$$
$$\xi_i > 0, i = 1,2,\ldots,K.$$

利用拉格朗日乘子法求解具有约束的二次规划问题,即:

$$\min\ Q(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\beta}) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{K} p_i\xi_i -$$
$$\sum_{i=1}^{K} a_i(y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) - 1 + \xi_i) - \sum_{i=1}^{K} \beta_i\xi_i,$$
$$\text{s.t. } a_i \geqslant 0, \beta_i \geqslant 0.$$

根据约束问题的 KKT 条件,有

$$\frac{\partial Q(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\beta})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{K} a_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i) = 0, \tag{3}$$
$$\frac{\partial Q(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\beta})}{\partial b} = -\sum_{i=1}^{K} a_i y_i = 0,$$
$$\frac{\partial Q(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\xi}} = p_i C - a_i - \beta_i = 0,$$
$$\frac{\partial Q(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\beta})}{\partial \boldsymbol{a}} = y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) - 1 + \xi_i = 0. \tag{4}$$

由式(3)得到

$$\boldsymbol{w}^* = \sum_{i=1}^{K} a_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i) = \sum_{i=1}^{S} a_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i).$$

式中:S 为拉格朗日乘子 $a_i > 0$ 所对应的样本(支持向量)总数量.根据式(4),能够求解出阈值 b^*,$(\boldsymbol{w}^*, b^*)$ 数值确定后,SVM 分类决策函数可由式(5)确定.

$$f(x) = \text{sign}(\boldsymbol{w}^* \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b^*) =$$
$$\text{sign}\left(\sum_{i=1}^{S} a_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b^*\right). \tag{5}$$

式中:$K(\boldsymbol{x}, \boldsymbol{x}_i)$ 为高维 Hilbert 空间中的内积.

3　实验与分析

3.1　数据集

为了评估基于加权聚类质心的 SVM 分类方法的性能,选择 6 组具有不同实际应用背景的 UCI 数据集进行测试.对于含有多个类别的数据,采用与其他文献相似的方法,即将其中的一类作为少数类,合并其他的类别为一个整体作为多数类.例如,将 page-blocks 的类别 5 作为少数类,合并其他的类作为多数类.实验数据集如表 1 所示.

表 1　UCI 数据集

Table 1　UCI Datasets

数据集	样例数目	少数类	多数类	不平衡度
vehicle	846	199	647	3.25
artificial	5 109	708	4 401	6.21
vowel	990	90	900	10.00
letter	20 000	734	19 266	26.25
abalone	4 177	115	4 062	35.32
page-blocks	5 473	115	5 358	46.59

3.2　评价标准

在传统的分类学习中,一般采用分类精度(分类正确的样本个数占总样本个数的百分比)作为评价指标,然而对于不平衡数据集,这一指标的实际意义不大,因为它反映的是多数类样本的分类测试结果.针对不平衡数据,很多学者提出了建立在混淆矩阵基础上的 F-measure、G-mean 等评价指标[11-12].

在某些应用中,人们更加关注少数类样本的分类性能,F-measure 就是用于衡量少数类分类性能的指标.F-measure 是查全率(recall)和查准率(precision)的调和均值,其取值接近两者的较小者,因此,较大 F-measure 值表示 recall 和 precision 都较大,其计算公式为

$$F_{\text{measure}} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

式中：$\text{recall}=\frac{\text{TP}}{\text{TP+FN}}$，$\text{precision}=\frac{\text{TP}}{\text{TP+FP}}$，TP、FN、FP、TN 的含义如表 2 所示.

表 2 2 类混淆矩阵

Table 2 A two-class confusion matrix

实际	预测	
	正类	反类
正类	TP	FN
反类	FP	TN

G-mean 是一种衡量数据集整体分类性能的评价指标，其值定义为

$$G_{\text{mean}}=\sqrt{A_{\text{positive}}\times A_{\text{negative}}}.$$

式中：正类分类准确率 $A_{\text{positive}}=\text{recall}=\frac{\text{TP}}{\text{TP+FN}}$，负类分类准确率 $A_{\text{negative}}=\frac{\text{TN}}{\text{TN+FP}}$.从定义中可以看出，G-mean 兼顾了少数类和多数类精度的平均，更能够反映出分类器的整体性能.

本文采用 F-measure 和 G-mean 作为分类方法的评价标准.

3.3 实验结果

为方便比较，实验中对数据采用五折交叉验证方式，为保证数据在进行分组过程中不平衡度保持一致，采用分层采样，将数据集中少数类和多数类样本分别随机分为 5 等份，两两随机组合得到 5 个与初始数据集不平衡度一致的子集，将其中 4 个子集作为训练集，1 个子集作为测试集，重复 5 次，以平均值作为最终的评价结果.

实验中，所提算法与其他 3 种具有代表性的方法进行比较：过采样 SMOTE 与决策树 C5.0 结合的 C5.0+SMOTE、随机欠采样与代价敏感支持向量机(CSVM)相结合的 undersample+CSVM、基于训练集划分与分类器集成的最小最大模块化支持向量机(M3-SVM)[13].实验结果如表 3 和 4 所示.

表 3 各种方法 F-measure 值比较

Table 3 The comparison of F-measure in different methods

数据集	C5.0+SMOTE	Undersample+CSVM	M3-SVM	本文算法
vehicle	0.850	0.613	0.869	**0.894**
artificial	0.643	0.516	0.682	**0.730**
vowel	0.698	0.493	**0.828**	0.766
letter	0.890	0.590	0.760	**0.907**
abalone	0.320	0.280	**0.392**	0.340
page-blocks	0.569	0.383	0.613	**0.680**
平均值	0.662	0.479	0.691	0.720

表 4 各种方法 G-mean 值比较

Table 4 The comparison of G-mean in different methods

数据集	C5.0+SMOTE	Undersample+CSVM	M3-SVM	本文算法
vehicle	0.810	0.612	0.815	**0.925**
artificial	0.713	0.654	0.753	**0.795**
vowel	0.787	0.769	**0.953**	0.939
letter	0.940	0.860	0.894	**0.992**
abalone	0.720	0.780	0.820	**0.940**
page-blocks	0.786	**0.896**	0.748	0.775
平均值	0.793	0.762	0.831	0.894

根据 F-measure 的定义，如果算法能够获得较大的值，则说明算法的少数类查全率和查准率都较高，有较好的少数类识别性能.从表 3 中可以看出，在所测试的 6 个数据集中，所提算法在其中 4 个数据集上的 F-measure 值都高于其他方法，特别是在 page-blocks 数据集上，与所比较 3 种算法中的最佳结果有将近 11%的提升.此外，通过计算 6 组 UCI 数据在不同方法下的 F-measure 平均值，也可以看出本文所提算法明显优于其他方法，在平均值性能上比次优算法 M3-SVM 有将近 5%的性能提升.

G-mean 的定义说明，其值越大，说明正类分类准确率和负类分类准确率都越大，也即算法在整体上的分类性能越好.从整体分类角度比较，本文算法在 6 组数据集上的 G-mean 度量指标值比次优算法有 8%的性能提升.

从 F-measure 和 G-mean 的实验结果中同时可以看出，在处理不平衡分类问题时，随机欠采样与代价敏感支持向量机相结合方法 undersample+CSVM 分类效果不佳.这是由于为了组成新的平衡训练集，随机欠采样方法丢弃了许多有用的负类样本信息，得到的分类超平面明显偏离实际位置，当不平衡度越高时，其劣势更加明显.此结果也间接表明本文所提出的通过以 K 均值聚类质心及其权重因子来代表众多负类样本的有效性，在组成新平衡训练数据集的同时，又不致于因改变数据集的数据分布而损失了负类样本所包含的有用的分类信息.

4 结束语

本文提出一种 K 均值聚类与支持向量机相结合的不平衡数据分类方法——WCC-SVM，该方法在进行数据预处理后，对训练集上的正负类样本各自进行无监督的 K 均值聚类，对每个聚类，以聚类质心和权重因子代表聚类样本分布和数量，然后进行基于加权 SVM 模型训练.实验结果表明，该方法在

显著降低实际参与模型训练样本数量的同时,能够取得不低于其他采样方法的分类性能,为大规模不平衡数据集分类问题提供了一种新的方法.

由于数据集本身的多样性和复杂性,样本的分布也呈现多样性,如果能估计正负类样本潜在的分布,根据不同的潜在分布设置不同的聚类方式,对算法的分类性能将会提高更多.

参考文献:

[1]叶志飞,文益民,吕宝粮.不平衡分类问题研究综述[J].智能系统学报, 2009, 4(2): 148-156.
YE Zhifei, WEN Yimin, LÜ Baoliang. A survey of imbalanced pattern classification problems[J]. CAAI Transactions on Intelligent Systems, 2009, 4(2):148-156.

[2]RONALDO C P, GUSTAVO E A, MARIA C M. A study with class imbalance and random sampling for a decision tree learning system[C]//International Conference for Information Processing. Milano, Italy, 2008: 131-140.

[3]WU Junjie, XIONG Hui, WU Peng, et al. Local decomposition for rare class analysis[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2007: 814-823.

[4]HE Haibo, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[5]李雄飞,李军,董元方,等.一种新的不平衡数据学习算法PCBoot[J].计算机学报, 2012, 35(2): 203-209.
LI Xiongfei, LI Jun, DONG Yuanfang, et al. A new learning algorithm for imbalanced data—PCBoost[J]. Chinese Journal of Computers, 2012, 35(2): 203-209.

[6]付忠良.不平衡多分类问题的连续 AdaBoost 算法研究[J].计算机研究与发展, 2011, 48(2): 2326-2333.
FU Zhongliang. Real AdaBoost algorithm for multi-class and imbalanced classification problems[J]. Journal of Computer Research and Development, 2011, 48(2): 2326-2333.

[7]VEROPOULOS K, CAMPBELL C, CRISTIANINI N. Controlling the sensitivity of support vector machines[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, USA, 1999: 55-60.

[8]AKBANI R, KWEK S, JAPKOWICZ N. Applying support vetor machines to imbalanced datasets[C]//Proceedings of 15th European Conference on Machine Learning. Pisa, Italy, 2004: 39-50.

[9]WU G, CHANG E Y. KBA: kernel boundary alignment considering imbalanced data distribution[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 786-795.

[10]ERTEKIN S, HUAN J, BOTTON L, et al. Learning on the border: active learning in imbalanced data classification[C]//Proceedings of the ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 127-136.

[11]HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//Proceedings of the International Conference on Intelligence Computing. Hefei, China, 2005: 878-887.

[12]DASKALAKI S, KOPANAS L. Evaluation of classifiers for an uneven class distribution problem[J]. Applied Artificial Intelligence, 2006, 20(5): 381-417.

[13]LÜ Biaoliang, WANG Kaian, UTIYAMA M, et al. A part-versus-part method for massively parallel training of support vector machines[C]//Proceedings of 17th International Joint Conference on Neural Networks. Budapest, Hungary, 2004, 1: 735-740.

作者简介:

胡小生,男,1978 年生,讲师,主要研究方向为机器学习、数据挖掘、信息检索.

钟勇,男,1970 年生,教授,博士,主要研究方向为信息检索、云计算.