DOI:10.16152/j.cnki.xdxbzr.2014.04.056

西北大学学报(自然科学版)
2014 年 8 月,第 44 卷第 4 期,Aug.,2014,Vol. 44,No. 4
Journal of Northwest University (Natural Science Edition)
JNWU

一种 LDA 和聚类融合的 SVM 多类分类方法

汝　佳,陈　莉,房鼎益

(西北大学 信息科学与技术学院,陕西 西安　710127)

摘要: 改进传统的基于二叉树结构的支持向量机多类分类方法。将无监督聚类引入到算法中,利用无监督聚类剔除大量的非支持向量样本,同时对于无监督聚类在异类样本相近时出现的性能下降问题,引入线性判别分析使得同类样本聚集,异类样本分散,确保聚类精度。线性判别分析和无监督聚类结合能够显著地缩减训练样本。该方法能够在保持分类准确率的情况下有效地提高 SVM 的分类速度。

关　键　词: 支持向量机; 线性判别分析; 模糊 C 均值聚类; 多类分类; 二叉树

中图分类号: TP391　　**文献标识码:** A　　**文章编号:** 1000-274X(2014)04-0559-04

SVM multi-class classification based on LDA and clustering

RU Jia, CHEN Li, FAN Ding-yi

(School of Information Science and Technology, Northwest University, Xi'an 710127, China)

Abstract: To improve traditional multi-class SVM method based on binary-tree. Using unsupervised clustering to extract training set, meanwhile, using linear discriminant analysis to solve the performance degradation of clustering when samples in different classes are similar, makes the samples in the same classes are gathered together and the samples in different classes are scattered, to ensure the accuracy of clustering. LDA and cluster can reduce training sample efficiently. The approach improves the speed of classification effectively while maintaining classification accuracy.

Key words: support vector machine; linear discriminant analysis; fuzzy C-means clustering; multi-class classification; binary-tree

Vapnik 提出的支持向量机 SVM(support vector machine)[1] 以训练误差作为优化问题的约束条件,以置信范围值最小化作为优化目标,即 SVM 是一种基于结构风险最小化准则的学习方法,其推广能力明显优于一些传统的学习方法。但是,SVM 是针对两类分类问题提出的,而实际中的多分类问题更为普遍,所以支持向量机多类分类方法的研究也就成为支持向量机研究的一个热点。常用的多类方法有:one-against-one(1-a-1)方法[2]、one-against-rest(1-a-r)方法[2]、有向无环图 DAG(directed acyclic graph)方法[4]、基于二叉树的方法[5]等。对于 N 类分类问题,1-a-1 方法需要构造 $N(N-1)/2$ 个两类 SVM 分类器,在分类的过程中采用投票法统计所有分类器的分类结果,选择的票数最多的类别作为待分类对象的类别,这种方法的缺点是需要训练的两分类分类器数目太多,训练速度慢;1-a-r 方法进行多类分类时虽然只需要训练 N 个分类器,但由于负类样本太多导致分类精度往往较低;DAG 方法也因需要的训练的 SVM 数目太多而训练速度较慢;基于二叉树的方法是近年来研究的热点,按结构可以分为"正二叉树"和"偏二叉树"两种[5],"正二叉树"是在内节点处,多个类和多个类的分割,"偏二叉树"则是一类样本与剩余类构造分割面。

收稿日期: 2013-04-13
基金项目: 国家自然科学基金资助项目(61070176)
作者简介: 汝佳,男,陕西渭南人,从事数据挖掘研究。

基于聚类思想的 SVM 多类分类方法也是目前研究的热点,针对该方面的研究已有很多[7-8]。本文引入聚类算法主要是提取可能含有支持向量的少部分样本作为训练样本[9],从而减少训练时间,提高分类效率,但是在每次聚类划分时经常会出现每个分组中均有多个类别的现象,造成训练样本缩减幅度小甚至没有缩减。针对以上问题,本文引入线性判别分析 LDA(linear discrimination analysis) 来优化特征,提高聚类精度,从而实现对算法的改进。

1 基于聚类的 SVM 多类分类

经典的 SVM 算法由于要进行大量的矩阵运算,在遇到大样本的情况下,需要占用很大的内存,分类速度会显著下降,所以需要减少训练样本来提高分类器的训练速度。由于支持向量机的分类面由支持向量决定,所以在减少训练样本时,防止降低支持向量机的分类性能,不能减少支持向量样本数目,需要减少的只是远离最优分类面处的训练样本。

模糊 C-均值聚类算法 FCM(fuzzy C-means clustering)[10] 实现简单,收敛速度快,所以本文利用它先将样本聚成 C 个子类,遍历每个子类,如果其中某个子类中的样本不属于同一个原始类别,则将该子类中的样本加入训练样本中,因为支持向量很可能存在于其中; 如果子类中样本都属于同一个原始类别,则将其加入测试样本,遍历结束后对测试样本进行分类,产生的错误样本加入训练样本中重新训练,直到错误样本数小于特定的值,这样保证了在不减少支持向量的前提下缩减了训练样本,提高了训练速度。

将此方法推广到多类分类中去,采用偏二叉树结构,选取一类作为正类样本,其余样本作为负类样本,逐层进行分类,直到只剩下一类为止。其中正类样本的选择采用类间最大距离法。设 $X = \{x_i, i = 1,2,\cdots,n\}$ 为待聚类的原始样本集,n 为样本个数,C 为待聚类数目,为各类的聚类中心,是第 k 个样本对第 i 类的隶属度,Y 为聚类后所选择的训练样本,则 FCM 的目标函数为

$$\min_{(U,V)}\{J_m(U,V) = \sum_{i=1}^{C}\sum_{k=1}^{n} u_{ik}^m,$$

$$D_{ik}^2\} D_{ik}^2 = \| x_k - v_i \|^2。$$

约束条件为: $\sum_{i=1}^{C} u_{ik} = 1(\forall k = 1,2,\cdots,n)$,其中加权指数 $m \geqslant 1$,训练详细步骤如下:

1) 指定聚类数目 C,初始化算法参数 m 和 v_i。C 的值和训练样本在空间中的分布有关。

2) 选取类间距离最大的类为正类,其余所有类为负类。即有 $d_{ij} = \min\{\| x_i - x_j \|\}$,其中 x_i 为第 i 类样本,取两类样本间欧氏距离的最小值为两类间距离,即 d_{ij},令 S_i 为第 i 类样本于其他所有类样本间距离的和,则选取 S_i 最大的一类为正类,其余所有类为负类。

3) 根据下式更新隶属度函数和聚类中心

$$u_{ik} = \left[\sum_{j=1}^{C}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}\right]^{-1},$$

$$\forall i,k \sim U_t = F_\partial(V_{t-1})$$

$$v_i = \left(\frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}\right), \forall i \sim V_t = G_\partial(U_{t-1})。$$

式中 t 表示第 t 次循环。

4) 当第 t 次更新聚类中心结束时,若有 $\| v_t - v_{t-1} \| \leqslant \varepsilon$,停止更新,原始样本被划分成 C 个子类。

5) 找出划分的 C 个子类中含有两类样本的子类,将其作为训练样本构建分类器。

6) 将所选正类样本从原始样本集中删除,重复上述步骤,若样本集中只含一类样本,算法停止,最终可以得到基于二叉树的 SVM 多类分类模型。

2 LDA 和聚类融合的 SVM 多类分类

2.1 线性判别分析

线性判别分析也叫作 Fisher 线性判别,是模式识别的经典算法,它是在 1996 年由 Belhunmeur 引入到模式识别和人工智能领域的[12]。线性判别分析的基本思想是将高维的模式样本投影到最佳的鉴别矢量空间,以达到抽取分类信息和压缩特征空间维数的效果,投影后保证模式样本在新的子空间有最大的类间聚类和最小的类内距离,即模式在该空间中具有最佳的可分离性。

假设对于一个 $\mathbf{R}^n$ 空间有 m 个样本分别为 x_1, x_2,$\cdots$,x_m,每个样本是一个 n 行的矩阵,其中 n_i 表示属于 i 类的样本个数,C 为样本类别数,则有:

i 类的样本均值为

$$u_i = \frac{1}{n_i} \sum_{i \in class\ i} x_i;$$

总体样本均值为

$$u_i = \frac{1}{m} \sum_{i=1}^{m} x_i;$$

类间离散度为

$$S_b = \sum_{i=1}^{C} n_i (u_i - u)(u_i - u)^{\mathrm{T}};$$

总类内离散度为

$$S_w = \sum_{i=1}^{C} \sum_{x_k \in class\ i} (u_i - x_k)(u_i - x_k)^{\mathrm{T}};$$

Fisher 判别准则表达式

$$J_{\mathrm{fisher}(\phi)} = \frac{\boldsymbol{\phi}^{\mathrm{T}} S_b \boldsymbol{\phi}}{\boldsymbol{\phi}^{\mathrm{T}} S_w \boldsymbol{\phi}}。$$

其中 $\boldsymbol{\phi}$ 为任意 n 维列矢量。Fisher 线性判别分析就是选取使得 $J_{\mathrm{fisher}(\phi)}$

达到最大值的矢量 $\boldsymbol{\phi}$ 作为投影方向，其物理意义就是使得投影后的样本具有最大的类间离散度和最小的类内离散度。

2.2　基于 LDA 和聚类的 SVM 多类分类

前面提到，在面临异类样本相近的情况下，聚类在每次划分时总会出现每个分组中均有多个类别的现象，因此会导致前面算法中的训练样本集根本没有缩减，算法的效率并没有提高。因此我们引入了线性判别分析优化样本特征，选择投影空间维数，找到最佳投影向量，使得同类样本紧缩，异类样本松散，从而很好地缓解了上述问题。

算法具体描述如下:

设 $X = \{x_i, i = 1, 2, \cdots, n\}$ 为原始待训练样本集，n 为样本个数，$Y = \{y_i, i = 1, 2, \cdots, n\}$ 为相应的样本标签。

1) 对当前待训练样本集 X 和样本标签 Y 执行 LDA，找到最佳投影向量 W 并保存，投影后样本为 X'。

2) 根据最大类间距离法找出相距其他类最远的那个类作为正类样本，为其余样本作为负类样本，重新建立只有两类的样本标签 Y'。

3) 对 X' 执行 FCM，将 X' 划分成 C 个子类，有 $X' = \{A_i, i = 1, 2, \cdots, C\}$，找出相应的样本标签 $Y' = \{B_i, i = 1, 2, \cdots, C\}$。

4) 找出满足包含两种或两种以上类别的子类 A_i，令 $S = S \cup A_i$，$L = L \cup B_i$，其中 S 和 L 初始为 $\varnothing$，剩余不满足条件的样本作为测试集 T。

5) 以 S 为训练样本，L 为样本标签训练 SVM 二值分类器，并对 T 进行分类，得到错误样本集 E，如果 E 的值小于规定的错误上限，则顺序执行下一步，否则令 $S = S \cup E$，重复本步。

6) 从 X' 和 Y' 中删去所选正类样本和相应的样本标签，如果 X' 中只有一类样本，算法结束，否则跳转至步骤(2)。

对于待识别样本，先用上述算法得到的投影向量 W 将其映射到特征空间，再根据正类的选择顺序送入相应的分类。

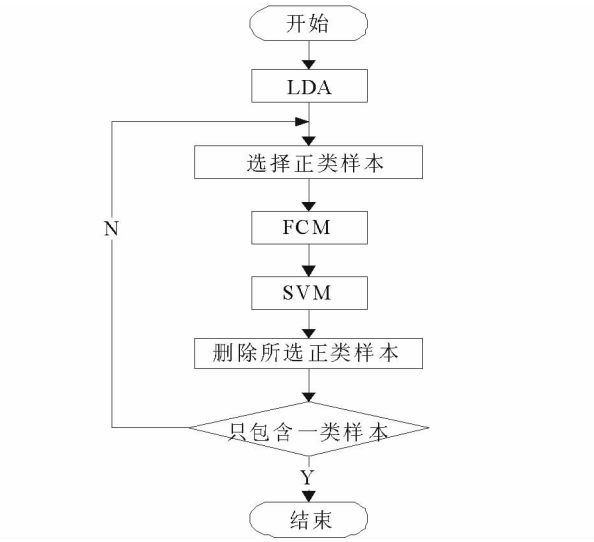

图 1　基于 LDA 和聚类思想的 SVM 多分类方法流程图

Fig. 1　Flow chart of SVM multi-class classification based on LDA and clustering

3　实　验

实验数据采用 UCI 数据集和土遗址健康状态分级数据，实验环境采用 Matlab R2010a，所用工具箱为台湾 Chih-Jen Lin 的 LIBSVM 工具箱和 FuzzyClust 聚类工具箱。

实验分为两部分，首先对比未引入 LDA 和引入 LDA 后两种多类分类 SVM 在相同数据集上选取的训练样本数目的差异，验证 LDA 对聚类的影响，其次在 UCI 数据集和一段时间内土遗址环境健康状态数据集上比较各分类算法的准确性，验证该算法的分类准确率。

偏二叉树结构下根据类别数 Iris，Wine，Glass 各需要训练 2 个、2 个和 5 个分类器，土遗址环境数据中则需要训练 4 个分类器。表 2 为 FCM-SVM 和经过 LDA 优化过的 FCM-SVM 两个算法在 Iris，Wine，Glass 和土遗址数据上每次构建分类器时选取的训练样本个数，表 3 为各多类分类

SVM方法在各个数据集上的准确率及消耗的时间,其中核函数选用径向基核函数。

表1 实验数据集

Tab.3 Experiment data set

	样本数	类别数	向量维数
Iris	150	3	4
Wine	178	3	12
Glass	214	6	9
Vowel	528	11	10
土遗址	5712	5	4

由表2可以看出,本文经过LDA优化的算法在训练样本的提炼上要稍优于FCM-SVM,在样本数较多的土遗址信息集中对训练样本的缩减尤为明显。表3中对比了每种方法的准确率和执行时间,其中以本文算法执行时间最短,并且在Iris和Glass数据集上的分类精度远远超出其他方法,在Wine和Vowel上基本持平,在土遗址数据中,本文算法同样表现突出。

表2 训练样本数对比结果

Tab.2 The number of training samples comparison

	FCM-SVM	本文算法
Iris	100/66	100/64
Wine	169/54	121/34
Glass	184/180/156/175/102	144/167/137/172/100
土遗址	4 900/4 914/4 509/418	4 348/596/100/679

表3 SVM多类分类实验结果对比

Tab.3 The comparion of SVM multi-class classification result

$C=1\ 000$, $\sigma=0.01$		$1-a-1$	$1-a-r$	DAG	本文算法
Iris	精度/%	97.33	96.67	96.67	99.99
	时间/s	0.143	0.316	0.107	0.010
Wine	精度/%	97.14	96.97	97.11	98.04
	时间/s	0.452	0.625	0.355	0.041
Glass	精度/%	75.89	72.10	76.63	97.30
	时间/s	1.213	5.173	1.501	0.746
Vowel	精度/%	97.81	97.73	97.58	96.33
	时间/s	0.915	6.415	1.546	0.246
土遗址	精度/%	95.03	87.35	91.68	93.74
	时间/s	4.301	14.269	2.223	1.035

4 结 语

本文对SVM多类分类方法进行研究,提出了基于聚类和偏二叉树结构的多类分类方法,并且引入LDA来优化样本特征,通过寻找最佳投影使同类样本聚集,异类样本分离,从而提高聚类精度,改善分类效果。实验结果表明,该方法在分类速度上明显优于常规多类分类算法,并且分类精度上也有所提高。但是,文中生成的二叉树不能很好地提高整个分类模型的推广能力,并且也未考虑FCM和SVM本身的优化及误差累积,这些也将作为进一步的研究目标,提升算法的执行效率。

参考文献:

[1] VAPNIK V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlage, 1995.

[2] HSU C, LIN C. A comparison of methods for multiclass support vector machines [J]. IEEE Trans on Neural Networks, 2002, 13(2): 415-425.

[3] TAKAHASHI F, ABE S. Decision-tree-based multiclass support vector machines [C] // Pro of 9th Int Confon Neural Information Processing, Singapore: [s. n.], 2002: 1418-1412.

[4] PLATT J, CRISTIANINI N, SHAWE-TAYLOR J. Large margin DAGs for multiclass classification [J]. Advances in Neural Information Processing Systems, 2000, 12(3): 547-553.

[5] 唐发明,王仲东,陈绵云. 支持向量机多类分类算法研究 [J]. 控制与决策,2005,20(7): 746-749.

[6] 唐发明,王仲东,陈绵云. 一种新的二叉树多类支持向量机算法 [J]. 计算机工程与应用,2005(7): 24-26.

[7] 陈增照,杨扬,何秀玲,等. 基于核聚类的SVM多类分类方法 [J]. 计算机应用,2007,27(1): 47-49.

[8] 赵晖,荣莉莉. 基于模糊核聚类的SVM多类分类方法 [J]. 系统工程与电子技术,2006,28(5): 770-774.

[9] 肖小玲,李腊元,张翔. 提高支持向量机训练速度的CM-SVM方法 [J]. 计算机工程与设计,2006, 27(22): 4183-4184.

[10] 赵炜,陈俊杰,李海芳. 融合LDA的多类SVM方法研究 [J]. 计算机工程与设计,2009,30(19): 4497-4499.

[11] 孙吉贵,刘杰,赵连宇. 聚类算法研究 [J]. 软件学报, 2008, 19(1): 48-61.

[12] BELHUMEUR P N, HESPANHA J P, KRIEGMAN D J. Eiegnfaces vs fisherfaces: Recognition using specific linear projection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.

(编 辑 曹大刚)