

一种新的基于二叉树的 SVM 多类分类方法

孟媛媛,刘希玉

(山东师范大学 信息管理学院,山东 济南 250014)

(myy801@126.com)

摘要:介绍了几种常用的支持向量机多类分类方法,分析其存在的问题及缺点。提出了一种基于二叉树的支持向量机多类分类方法(BT-SVM),并将基于核的自组织映射引入进行聚类。结果表明,采用该方法进行多类分类比 1-v-r SVMs 和 1-v-1 SVMs 具有更高的分类精度。

关键词:多类分类;支持向量机;二叉树;自组织映射

中图分类号:TP181;TP391 **文献标识码:**A

A new SVM multiclass classification based on binary tree

MENG Yuan-yuan, LIU Xi-yu

(College of Information and Management, Shandong Normal University, Jinan Shandong 250014, China)

Abstract: The problems and defections of the existing methods of SVM multi-class classification were analyzed. A multi-class classification based on binary tree was put forward. A modified self-organization map (SOM), KSOM (kernel-based SOM), was introduced to convert the multi-class problem into binary tress, in which the binary decisions were made by SVMs. The results show that it has higher multiclass classification accuracy than the multi-class SVM approaches with "one-versus-one" and "one-versus-the rest".

Key words: multi-class classification; support vector machine(SVM); binary tree; Self-Organization Map(SOM)

支持向量机(Support Vector Machine, SVM)是近几年发展起来的一种学习机器^[1],是结构风险最小化方法的近似实现。通过学习, SVM 可以自动寻找那些对分类有较好区分能力的支持向量,由此构造出的分类器可以最大化类之间的间隔,因而具有较好的推广能力和较高的分类准确率,能用于模式分类和非线性回归。但是传统的 SVM 是针对两类分类问题设计的,不能直接用于多类分类问题。本文提出了一种基于二叉树的 SVM 多类分类方法,通过先聚类后分类的思想构造最优二叉树结构,以提高分类精度,并取得了较为理想的结果。

1 常用多类支持向量机方法

支持向量机方法最初是针对两类分类问题而提出的,如何将两类分类方法扩展到多类别分类是支持向量机研究的重要内容之一。 K 分类问题($K > 2$)和两类分类问题之间存在一定的对应关系,如果一个分类问题 K 可分,则这 K 类中的任何两类间一定可分;反之,在一个 K 分类问题中,如果已知任意两两可分,则通过一定的组合法则,可由两两可分来最终实现 K 类可分。

假定多类分类问题有 K 个类别 $S = \{1, 2, \dots, k\}$, 训练样本为 $\{(x_i, y_i), i = 1, 2, \dots, l\}$, 其中 $y_i \in S$ 。目前有以下一些常用方法实现支持向量机的多类别分类。

1.1 1-v-r SVMs

1-v-r 方法(One-versus-the-rest Method)^[2]构造 k 个支持向量机子分类器。在构造第 i 个支持向量机子分类器时,将属于第 i 类别的样本数据标记为正类,不属于 i 类别的样本数据标记为负类。测试时,对测试数据分别计算各个分类器的决策

函数值,并选取函数值最大对应的类别为测试数据的类别。

1.2 1-v-1 SVMs

1-v-1 方法(One-versus-one Method)^[2]是由 Knerr 提出的,该算法在 k 类训练样本中构造所有可能的两类分类器,每类仅仅在 k 类中的 2 类训练样本上训练,结果共构造 $k(k-1)/2$ 个 SVM 子分类器。在构造类 i 和类 j 的 SVM 子分类器时,在样本数据集选取属于类 i 和类 j 的样本数据作为训练样本数据,并将属于类 i 的数据标记为正,将属于类 j 的数据标记为负。测试时,将测试数据对 $k(k-1)/2$ 个 SVM 子分类器分别进行测试,并累计各类的得分,选择得分最高者所对应的类为测试数据的类别。如图 1,图 2 所示,1-v-r 和 1-v-1 方法都存在不可区分区域(阴影部分)。

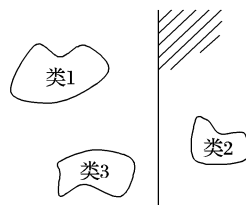


图 1 1-v-r SVMs

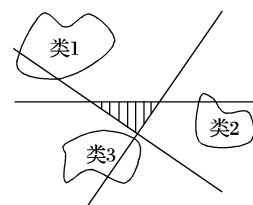


图 2 1-v-1 SVMs

1.3 DDAG SVMs 方法

DDAG SVMs 方法是 Platt 等提出的决策导向非循环图(Decision Directed Acyclic Graph, DDAG)^[3]方法,将多个两类分类器组合成多类分类器。在训练阶段与 1-v-1 方法相同,对 k 类问题, DDAG 含有 $k(k-1)/2$ 个两类分类器;然而在决策阶段,使用从根结点开始的导向非循环图(DAG),具有 $k(k-1)/2$ 个内部结点以及 k 个叶子结点,每个内部结点都是一

收稿日期:2005-05-13;修订日期:2005-07-20

基金项目:山东省自然科学基金资助项目(Z2004G02);山东省中青年科学家奖励基金项目(03BS003)

作者简介:孟媛媛(1980-),女,山东临沂人,硕士研究生,主要研究方向:人工神经网络、模式识别;刘希玉(1965-),男,山东莱芜人,教授,博士生导师,主要研究方向:非线性模型理论、神经网络遗传算法。

个两类分类器,叶子结点为最终的类值。如图3所示。给定一个测试样本,从根结点开始根据分类器的输出值决定其走左侧或右侧路径,如此一直到叶子结点为止,得到样本所属的类值。其优点是推广误差只取决于类数 k 和结点上的类间间隙(Margin),而与输入空间的维数无关。

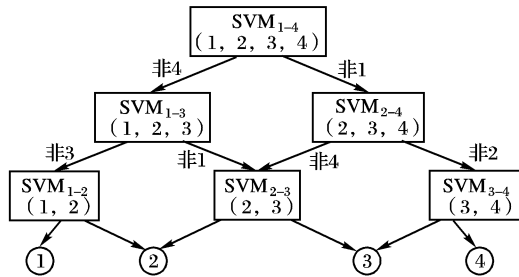


图3 DDAGSVMs 分类

2 基于二叉树的多类支持向量机分类

以上三种多类分类方法都存在一些不足,如存在不可分盲点,或需要训练的支持向量机个数太多,或分类未知样本时使用的支持向量机过多。因此本文提出了一种将支持向量机和二叉树的基本思想相结合的多类分类方法(BT-SVM),以获得最优的二叉树结构。

2.1 构造合理的二叉树结构

二叉树结构分类器可以把一个复杂的多类别问题化为多个两类问题来解决。一个多类别分类问题转化为两类问题的形式是多种多样的,对应的二叉树的结构也各不相同。对于一个 K 类问题,所有可能构造的严格二叉树的数目为 $N_k =$

$\prod_{i=1}^{k-1} 2 * i - 1$,其中 $k > 1$ ^[4]。不同的层次结构对分类精度有一定影响,并且这种影响有可能产生“误差累积”现象^[5],即若在某个结点上发生分类错误,将会把错误延续下去,该结点后续下一级结点上的分类就失去意义。由以上的分析可知,设计一个二叉决策树需要选择一个合适的树结构,即合理安排树的结点和分支。为使二叉决策数具有最优的性能,必须在决策结点,以近似最优的方法将多类样本分为两组,使两组样本的聚类中心距离最大,且每组样本数据分歧最小,即使上层中两个子类之间的可分性尽可能强,以构造合理的层次结构。

采用先聚类再分类的思想,即通过聚类,将 K 类样本聚类为两类,由此将 K 类问题转化为两类分类问题,然后用转化后的两类样本构造该决策结点的最优超平面,以构造合理的二叉树结构。

采用先聚类再分类的思想,即通过聚类,将 K 类样本聚类为两类,由此将 K 类问题转化为两类分类问题,然后用转化后的两类样本构造该决策结点的最优超平面,以构造合理的二叉树结构。

2.2 聚类及分类

2.2.1 基于核函数的自组织映射(KSOM)

自组织映射网络^[6]是一种竞争式神经网络,具有聚类、自组织、自学习以及可视化的功能。传统SOM网络是采用欧氏距离作为评价神经元竞争获胜的标准的,它根据样本的内在联系,对样本自动进行聚类。由于支持向量机分类器是基于核空间的,故我们引入一种基于核函数的SOM方法,与传统SOM方法不同的是:其度量样本和权向量之间的尺度不再使用欧氏距离,而是使用核代入方法。

传统的SOM算法中,在时刻 $n+1$,神经元 j 的权值向量被定义为:

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n)h_{ji}(n)(\mathbf{x}(n) - \mathbf{w}_j(n)), \quad \mathbf{x} \in R^N \quad (1)$$

基于核方法,可通过由核诱导的非线性映射 Φ 将输入空

间的非线性问题变换至高维(甚至无穷维)特征空间中较易解决的线性问题,引入后得到新的调整公式:

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n)h_{ji}(n)(\Phi(\mathbf{x}(n)) - \mathbf{w}_j(n)), \quad \mathbf{w}_j, \Phi \in R^M \quad (2)$$

其中 $N \ll M$, \mathbf{w}_j 为神经元 j 的突触权值向量; i 为获胜神经元, $\Phi(\mathbf{x})$ 为 \mathbf{x} 在特征空间诱导的像, h_{ji} 为获胜神经元 $i(\mathbf{x})$ 周围的邻域函数; $\eta(n)$ 是学习参数。

$$\text{我们可以将 } \mathbf{w}_j \text{ 定义为: } \mathbf{w}_j(n) = \sum_{k=1}^L a_{jk}^{(n)} \Phi(\mathbf{x}_k)$$

将其代入到公式(2),可得:

$$\begin{aligned} & \sum_{k=1}^L a_{jk}^{(n+1)} \Phi(\mathbf{x}_k) g\Phi(\mathbf{x}_i) \\ &= (1 - \eta(n)h_{ji}(n)) \sum_{k=1}^L a_{jk}^{(n)} \Phi(\mathbf{x}_k) g\Phi(\mathbf{x}_i) + \\ & \eta(n)h_{ji}(n) \Phi(\mathbf{x}_i) g\Phi(\mathbf{x}_i) \end{aligned} \quad (3)$$

其另一种表达公式为:

$$a_j^{(n+1)} K = (1 - \eta(n)h_{ji}(n)) a_j^{(n)} K + \eta(n)h_{ji}(n) k_i \quad (4)$$

其中 $K = \{k(\mathbf{x}_j, \mathbf{x}_i)\}_{(j,i)=1}^L$; $k(\mathbf{x}_j, \mathbf{x}_i) = \Phi(\mathbf{x}_j) g\Phi(\mathbf{x}_i)$; $a_j^{(n)} = [a_{j1}^{(n)}, a_{j2}^{(n)}, \dots, a_{jL}^{(n)}]^T$; $k_i = [k(\mathbf{x}_1, \mathbf{x}_i), k(\mathbf{x}_2, \mathbf{x}_i), \dots, k(\mathbf{x}_L, \mathbf{x}_i)]$

由此对于获胜神经元 i ,基于KSOM的调整规则变为:

$$A_{ji}^{(n+1)} K = (1 - \eta(n)h_{ji}(n)) A_{ji}^{(n)} + \eta(n)h_{ji}(n) \quad (5)$$

对于输入向量 \mathbf{x} ,确定获胜神经元 i 的公式为:

$$i(\mathbf{x}) = \arg \min_j \|\Phi(\mathbf{x}) - \mathbf{w}_j(n)\| \quad (6)$$

$$\text{由 } \|\Phi(\mathbf{x}) - \mathbf{w}_j(n)\|^2 =$$

$$k(\mathbf{x}, \mathbf{x}) + A_j^{(n)} K^T \{A_j^{(n)}\}^T - 2 \sum_{k=1}^L a_{jk}^{(n)} k(\mathbf{x}_k, \mathbf{x}) \quad (7)$$

代入公式(6),可得对于输入向量 \mathbf{x}_i ,获胜的神经元为:

$$i(\mathbf{x}) = \arg \min_j \{k(\mathbf{x}_i, \mathbf{x}_i) + A_j^{(n)} K^T \{A_j^{(n)}\}^T - 2 A_j^{(n)} k_i^T\}$$

通过基于核函数的自组织映射(KSOM)进行聚类,可以将 K 类样本聚类成两个子集。

2.2.2 分类过程

通过逐层聚类,即可确定二叉决策树的结构,从而确定每个决策结点的最优分解面。每个决策结点是用一个支持向量机实现的,使用支持向量机可以保证在既定的分类器结构下,单个决策结点的分类间隔最大。样本分类的具体过程如下:

1) 令初始状态只含有根结点(X),其中 X 为全体训练样本的集合;

2) 若所有新结点(X')只含有一个类,则用这些新结点所对应的类标志这些新结点为叶结点,学习算法结束;

3) 否则,标志该新结点为决策结点;应用上述聚类算法将该决策结点的训练样本集 X' 聚类成两个子集 X_{+1}' 和 X_{-1}' ;用支持向量机的学习算法求得该决策结点的最优分类面;

4) 由 X_{+1}' 和 X_{-1}' 形成两个新结点:(X_{+1}')和(X_{-1}');转步骤2)继续学习。

2.3 训练时间分析

支持向量机的训练时间与样本集的大小成超线性关系^[7], $T = ck^\alpha$,其中 c 为比例常数, α 为一常数,大小与不同的分解算法有关,对于基于分解方法的支持向量机学习算法来说, $\alpha \approx 2$,因此SVM的训练时间取决于参与训练的样本的数量多少。

设样本集大小为 k ,模式类别数为 K ,对于标准的1-v-r方法,需要训练 K 个支持向量机,则总的训练时间为: $T_{1-v-r} = cKk^\alpha$ 。

(下转第2657页)

$$\leq |x_{n1}' - x_{n2}'| + |x_{(n+1)1}' - x_{(n+1)2}'|$$

所以, $AF(A)$ 和 $AF(A')$ 的分子有下列关系:

$$\sum_{i=1}^{n+1} |x_{i1}' - x_{i2}'| - \sum_{i=1}^n |x_{i1} - x_{i2}| \geq 0$$

但由于 $AF(A)$ 和 $AF(A')$ 的分母上分别为 n 和 $n+1$, 所以 $AF(A)$ 和 $AF(A')$ 没有恒定的大小关系。故 AF 算法从根本上避免了多值偏向问题。

3 AF 算法的评估

表 2 隐形眼镜数据集实验结果

训练集	测试集	AF 分类正确率	ID3 分类正确率
12	12	0	0.08
16	8	0.75	0.75
18	6	1.0	1.0
19	5	1.0	0.8
20	4	0.75	0.75
21	3	1.0	1.0
22	2	1.0	1.0

表 3 tic-tac-toe 数据集实验结果

训练集	测试集	AF 分类正确率	ID3 分类正确率
479	479	0.81	0.80
638	320	0.86	0.88
718	240	0.88	0.85
766	192	0.86	0.81
822	137	0.83	0.83
862	96	0.83	0.84
910	48	0.81	0.85

(上接第 2654 页)

对于 1-v-1 SVMs 和 DDAG SVMs 方法, 需要训练 $K(K-1)/2$ 个支持向量机, 假设每个模式类所包含的训练样本数目相等, 则每个支持向量机的训练样本数为 $2k/K$, 总的训练时间为: $T_{1-v-r} = T_{DDAG} = c \frac{K(K-1)}{2} \left(\frac{2k}{K}\right)^\alpha \approx 2^{\alpha-1} c K^{2-\alpha} k^\alpha$ 。

当 $\alpha = 2$ 时, 两种方法的训练时间与类别数无关。

对于 BT-SVM 方法, 一般情况下, $k \gg K$, 故分类器构造时间会远远少于支持向量机的学习时间, 因此分类器构造时间忽略不计。假设分类器的结构为完全二叉树结构, 每个类包含的训练样本数目相等, 支持向量机的个数为 $K-1$, 高度 $H(i)$ 的决策结点的支持向量机的训练样本数为 $k/2H(i) - 1$ 。其中, $H(i)$ 为叶结点的高度。总的训练时间为: $T_{BT-SVM} = \left(\frac{1 - 2^{\log_2^K(1-\alpha)}}{1 - 2^{1-\alpha}}\right) c k^\alpha$ 。

从训练时间来看, 由于二叉树的多类分类支持向量机个数明显少于前三种, 因此训练时间更短, BT-SVM 方法要优于其他三种多类分类方法。

3 实验及结果分析

通过在 UCI 数据库的 wine, iris, vowel, glass 四个数据上进行实验, 将本文提出的 BT-SVM 方法与 1-v-r SVMs 和 1-v-1 SVMs 进行比较。使用径向基函数作为核函数, 结果如表 1 所示。

从实验结果可以看出, BT-SVM 与其他几种分类器相比, 整体性能和分类精度大大提高。在对实验结论进行的统计显著性分析来看, 该 BT-SVM 的分类正确性明显高于其他几种

下面我们将对 AF 算法的分类正确率作出评估, 并与 ID3 算法的分类正确率进行比较。使用两个数据集。第一个数据集是上文所用到的隐形眼镜数据集, 另外一个 tic-tac-toe 游戏的数据集。

隐形眼镜数据集^[5]共有 24 个数据, 4 个普通属性, 1 个分类属性。普通属性及其取值分别为: 年龄(青少年、中年、老年), 眼疾(近视、远视), 散光(有、无), 眼泪(减少、正常)。类别属性取值为:(适合、不适合)。

tic-tac-toe 游戏数据集^[5]共有 958 个数据, 9 个普通属性, 1 个分类属性。9 个普通属性分别代表游戏面板上的 9 个位置, 取值均为(x, b, o), x 表示在该位置放置一个 x, b 表示该位置是空的, o 表示在该位置放置一个 o。分类属性的取值为(positive, negative), 代表 x 一方的输赢状况, positive 表示 x 一方赢, negative 表示 x 一方输。如果首先出现 3 个 x 位于同一行、同一列或者同一对角线则 x 一方赢, 如果首先出现 3 个 o 位于同一行、同一列或者同一对角线则 o 一方赢。

实验结果发现, AF 算法和 ID3 算法在两个数据集上的分类正确率差别不大, 各有优劣。

参考文献:

- [1] QUINLAN JR. Induction of decision tree [J]. Machine Learning, 1986, (1): 81-106.
- [2] 孙毅. 数据挖掘中的决策树方法及其在客户分类中的应用[D]. 大连: 大连理工大学, 2004.
- [3] 曲开社, 成文丽, 王俊红. ID3 算法的一种改进算法 [J]. 计算机工程与应用, 2003, (25): 104-107.
- [4] QUINLAN JR. C4.5: Programs for Machine Learning [M]. San Mateo, CA: Morgan Kaufmann, 1993.
- [5] University of California Irvine. UCI KDD Archive[DB/OL]. <http://kdd.ics.uci.edu/>, 2005-03-21.

多类分类方法。

表 1 几种多类分类方法的分类精度比较

数据	参数 σ	分类精度(%)		
		1-v-1 SVMs	1-v-r SVMs	BT-SVM
wine	3	97.77	97.76	98.19
iris	1	97.02	97.02	97.09
vowel	5	97.35	96.66	97.59
glass	0.3	83.35	81.56	85.18

参考文献:

- [1] VAPNIK V. The Nature of Statistical Learning [M]. New York: Springer Verlag, 1995.
- [2] HSU C-W, LIN C-J. A comparison of methods for multi-class support vector machines [J]. IEEE Transaction on Neural Network, 2002, 13(2): 415-425.
- [3] PLATT JC, CRISTIANINI N, SHAWE-YAYLOR J. Large Margin DAGs for multiclass classification [A]. Advances in Neural Information Processing Systems [C], 2000. 547-553.
- [4] BEILEY A. Class-dependent features and multicategory classification [D]. navy.mil/csf/papers/baileyphd.pdf, 2001.
- [5] 刘志刚, 李德仁, 秦前清, 等. 支持向量机在多类分类问题中的推广 [J]. 计算机工程与应用, 2004, 40(7): 10-13.
- [6] HAYKIN S. Neural Networks: A Comprehensive Foundation [M]. 2nd Edition. 叶世伟, 史忠植, 译. 北京: 机械工业出版社, 2004. 321-347.
- [7] PLATT JC. Fast Training of Support Vector Machines using Sequential Minimal Optimization [A]. Advances in Kernel Methods: Support Vector Learning [C]. Cambridge: MIT Press, 1999. 185-208.