Journal of Computer Applications
计算机应用,2013,33(10):2801 -2803
ISSN 1001-9081
CODEN JYIIDU
2013-10-01
http://www.joca.cn

文章编号:1001-9081(2013)10-2801-03
doi:10.11772/j.issn.1001-9081.2013.10.2801

基于分层聚类及重采样的大规模数据分类

张　永*,浮盼盼,张玉婷
(辽宁师范大学 计算机与信息技术学院,辽宁 大连 116081)
(*通信作者电子邮箱 zhyong@lnnu.edu.cn)

摘　要:针对大规模数据的分类问题,将监督学习与无监督学习结合起来,提出了一种基于分层聚类和重采样技术的支持向量机(SVM)分类方法。该方法首先利用无监督学习算法中的*k*-means聚类分析技术将数据集划分成不同的子集,然后对各个子集进行逐类聚类,分别选出各类中心邻域内的样本点,构成最终的训练集,最后利用支持向量机对所选择的最具代表样本点进行训练建模。实验表明,所提方法可以大幅度降低支持向量机的学习代价,其分类精度比随机欠采样更优,而且可以达到采用完整数据集训练所得的结果。

关键词:海量数据; 分类; 聚类; 重采样; 支持向量机
中图分类号: TP181　　**文献标志码**:A

Large-scale data classification based on hierarchical clustering and re-sampling

ZHANG Yong*, FU Panpan, ZHANG Yuting
(*School of Computer and Information Technology, Liaoning Normal University, Dalian Liaoning* 116081, *China*)

Abstract: Based on hierarchical clustering and re-sampling, this paper presented a Support Vector Machine (SVM) classification method for large-scale data, which combined supervised learning with unsupervised learning. The proposed method first used *k*-means cluster analytical technology to partition dataset into several subsets. Then, the method clustered class by class for each subset and selected samples in each clustering center neighborhood to form candidate training datasets. Last, the method applied SVM to train and model for candidate training datasets. The experimental results show that the proposed method can substantially reduce SVM learning cost. Meanwhile, the proposed method has better classification accuracy than random re-sampling method, and can attain about the same classification accuracy of the non-sampling method.

Key words: large-scale data; classification; clustering; re-sampling; Support Vector Machine (SVM)

0　引言

由于较高的分类精度和良好的泛化能力,基于统计学习理论的支持向量机(Support Vector Machine, SVM)是最有效的分类方法之一[1]。然而,近年来随着社会网络的发展与计算机技术的不断进步,人们能够获得的信息量与日俱增,比如各种大型视频库、图像库、语料库等,其数据规模越来越大。要分类规模如此庞大的数据集,支持向量机在学习过程中需要占用大量内存,寻优速度非常缓慢,因此支持向量机对大规模数据集训练速度慢的瓶颈凸显出来[2]。

为此,学者们进行了大量的研究,试图解决支持向量机对大规模数据集训练速度慢的瓶颈,其方法大致可以分为两类:一类是改进SVM求解算法,比如,Chen等[3]通过引入分而治之的思想,采用序列最小优化算法将SVM的二次规划问题分解为多个子问题,以期提高SVM的训练速度;Huang等[4]结合人工神经网络算法来简化SVM的训练过程;Dong等[5]引入了并行优化步骤,用块对角矩阵近似代替原始的核矩阵,从而加快SVM训练速度。这些方法在某种程度上确实提高了样本训练速度,但对于大规模数据集依然不是很理想。另一类方法是借助一些其他算法来缩减数据规模,约简训练集,比如基于随机采样的SVM算法、基于聚类的SVM算法[6-8];Cervantes等[9]将聚类与SVM相结合,计算每一类的质心并作为此类的代表样本点,提出了基于最小内附球的SVM快速训练算法;陈光喜等[10]针对SVM对大规模数据集训练速度慢的瓶颈,提出了一种聚簇消减数据集方法。但这些方法在缩减数据规模的同时,忽略了数据样本本身的分布特性,可能导致被选择的分类样本信息含量减少,从而影响分类精度。

本文采用约简训练集的算法思想,提出一种新的学习策略:首先利用*k*-means聚类算法将大数据集划分成*K*个规模减小了的子集,对于各个子集,基于其分布密度,按照某种规则对各个子集块中的每一类进行模糊C均值(Fuzzy C-Means, FCM)聚类,选取出类中心邻域内的样本点进行SVM训练建模。该方法在遵循减少样本点数量的前提下,最大限度地保证了训练数据集的信息含量,同时还加入了原数据样本点的分布信息,将各类样本在原数据分布中的重要性考虑进来。实验结果表明,该方法在保证分类精度的基础上提高了SVM的分类速度。

1　基于分层聚类及重采样的数据分类方法

1.1　基于分层聚类与重采样的样本选取策略

聚类是无监督学习算法的典型代表之一,它能够按照一定的要求和规律对数据集进行区分,把一个没有类别标记的数据集划分成若干个子集,使相似的数据尽可能地划分到同一子集中。聚类在大规模数据约简中具有广泛的应用。

收稿日期:2013-03-13;**修回日期**:2013-05-29。　**基金项目**:国家自然科学基金资助项目(61373127);中国博士后科学基金资助项目(20110491530);辽宁省教育厅基金资助项目(L2011186)。

作者简介:张永(1975 -),男,四川阆中人,副教授,博士,CCF会员,主要研究方向:机器学习、智能计算;　浮盼盼(1987 -),女,河南新乡人,硕士研究生,主要研究方向:机器学习;　张玉婷(1990 -),女,黑龙江哈尔滨人,硕士研究生,主要研究方向:机器学习。

为了约简大规模数据集,同时保证数据集的信息含量,即挑选出最具代表性的训练样本点,本文提出了基于分层聚类与重采样的样本选取方法。基本思想包括两部分:首先,利用 k-means 算法将数据集 D 划分成 K 块,每一块都是 D 的子集,即有 $D = D_1 \cup D_2 \cup \cdots \cup D_K$,且 $D_i \cap D_j = \varnothing\ (i, j = 1,2,\cdots, K, i \neq j)$。其次,对于每个子集 D_i,根据其样本分布特性,采用基于密度的 FCM 算法,对其所含的每一类样本(不妨设为第 j 类)进行聚类,选出各类中心邻域内的样本点作为第 i 个子集块 D_i 中第 j 类的候选训练样本,并将其加入到新的训练集中。

本文提出的样本选取策略的关键在第二步,即采用基于密度的 FCM 算法来选取适当的样本进行训练。为了更好地选取样本,本文需要解决两个问题:其一,确定每个子类中选取的样本数;其二,确定类中心邻域。首先给出类中心邻域的定义。

定义 1 类中心邻域。对于给定的某个样本类,类中心标记为 v,*mean_dist* 表示类间样本平均距离,则以 v 为中心,*mean_dist* 为半径的圆形区域,称为类中心邻域。落在其邻域内的样本点的数目称为类中心密度,记为 *density*(v)。*density*(v)计算如下:

$$density(v) = \sum_{i=1}^{n} u(mean\_dist - d(x_i, v)); \quad u(k) = \begin{cases} 1, & k \geqslant 0 \\ 0, & k < 0 \end{cases} \tag{1}$$

其中 $d(x_i, v)$ 为类中心 v 与样本 x_i 的距离,n 为类中样本数。

明确了类中心邻域后,算法还需确定每个子类中选取的样本数。显然,为了保证提出方法的有效性,所选取的样本不仅要包含丰富的信息,而且还要尽量不影响数据的分布特性。为此,给出了第 i 个子集 D_i 中第 j 类应选取的样本数 l_{ij} 为

$$l_{ij} = (n_t / n) \times m_{ij} \tag{2}$$

其中:n 为数据集 D 的样本总数,n_t 为预选取的训练样本数,m_{ij} 为第 i 个子集 D_i 中第 j 类的样本数目。在对子集 D_i 的第 j 类进行 FCM 聚类时,如果样本数目足够多,则以式(2)计算所得的 l_{ij} 为聚类数。显然,这个聚类数是根据样本分布特性自适应选取的。在第 j 类进行 FCM 聚类后,将形成 l_{ij} 个聚类中心,把每个聚类中心邻域内的 *density*(v)个样本点都加入到候选训练集中。

图 1 和图 2 以 wine 数据集为例,对比说明了上述处理前后的数据分布情况。利用本文方法从 wine 数据集中选取了 60 个样本。从图 2 可以看出,经过本文的选取方法,不仅有效地减少了训练样本的数量,同时还较好地保持了数据集的原始分布信息。

1.2 算法描述

给定一个包含 n 个样本的数据集 D,根据上述的思想,本文提出了一个基于分层聚类与重采样的 SVM 分类算法,具体描述如算法 1 所示。

算法 1 基于分层聚类与重采样的 SVM 分类。

输入 数据集 D,预选取的训练样本数 N,k-means 的聚类个数 K。

输出 SVM 分类模型,整体分类精度 r。

步骤 1 利用 k-means 算法将数据集 D 划分为 K 个子集块,使得 $D = D_1 \cup D_2 \cup \cdots \cup D_K$,初始训练集 *Training_set* = { }。

步骤 2 利用基于密度的 FCM 算法对每一个子集块 D_i 进行逐类聚类,得到训练子集 T_i。具体方法为:对每个子集块 D_i,设 T_i 初始值为空,采用基于密度的 FCM 聚类算法对其中的每一类 D_{ij}($j = 1,2,\cdots,s$,其中 s 为该子集块中所包含的类别数) 进行聚类:

1)利用式(2)计算应预选取的样本数 l_{ij};

2)如果 l_{ij} 小于某个阈值,则将该子类样本全部加入训练子集 T_i;否则以 l_{ij} 为聚类数进行 FCM 聚类,得到类中心 v_k,并利用式(1)计算得到类中心密度 *density*(v_k),将 v_k 邻域内的 *density*(v_k)个样本点加入训练子集 T_i;

3)把得到的训练子集 T_i 加入最终的训练集 *Training_set* = *Training_set* $\cup$ T_i。

步骤 3 利用得到的训练集 *Training_set* 进行 SVM 训练;

步骤 4 使用步骤 3 中建立的模型对测试集进行测试,并计算分类精度 r。

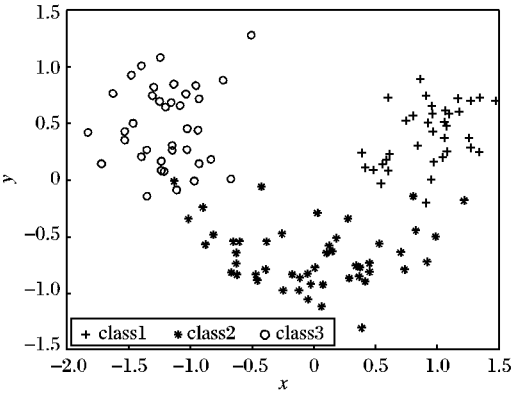

图 1 wine 数据集的原始分布情况

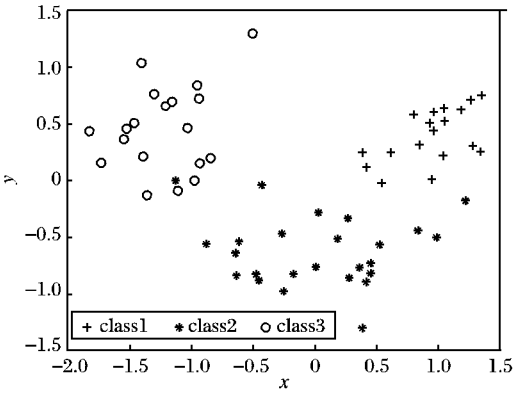

图 2 选取 60 个代表样本的分布情况

2 实验结果与分析

2.1 实验环境及数据集

为了分析提出的样本选择策略的有效性,将本文方法与随机采样和无采样进行了实验对比,分类方法都基于 LIBSVM 工具箱[11]实现。在实验中,对 SVM 的参数 c 和 g 使用网格搜索法进行寻优,寻优范围设置为$[2^{-10}, 2^{10}]$。为了实验对比,三个算法均通过 5 折交叉验证方法获得其均值,且事先为每个数据集选用了相同的测试集。对于分层聚类中的 FCM 算法,本文将其模糊因子设置为 2。

实验中所使用的 6 个数据集都来自于 UCI 数据库[12],其数据信息如表 1 所示。其中,对于 covtype 数据集,鉴于其不平衡率较高,本文只选取了第 3、5、6、7 类中的 83 124 个样本来进行实验(本文用 covtype* 表示)。

表 1　数据信息表

数据集	样本数	特征数	类数
wine	178	13	3
pima	768	8	2
statlog	6435	36	6
letter	20000	16	28
adult	30138	14	2
covtype*	83124	55	7

2.2　实验结果与分析

首先给出了本文方法的一些实验结果,如表 2 所示,分别给出了预选取的训练样本数,经过本文算法后实际选取的训练样本数,在获取最好分类精度时 k-means 划分的块数、支持向量数和分类精度。其次,将本文方法与随机采样和无采样进行了实验对比,实验结果如表 3 所示。其中随机采样方法是按照一定的比例从大规模数据集中随机选取样本,然后再用SVM进行训练和测试,称为方法一;无采样方法是指不对原始训练样本进行采样,而直接用 SVM 进行训练和测试,称为方法二;本文方法先基于分层聚类重采样,然后再用 SVM 进行训练和测试。表 3 给出了三种方法的支持向量数、分类精度和训练时间三个指标。对于分层聚类重采样,本文选取 k-means 划分不同块数实验结果的平均值,对于 letter、adult、covtype* 三个大规模数据集,采用完整训练集直接进行 SVM 训练时,时间代价远远大于另外两种方法,这里不列入对比,以“—”表示。

表 2　训练样本数对比

数据集	预选训练样本数	实际训练样本数	k-means 划分块数	支持向量数	分类精度/%
wine	60	63	5	56	100.00
pima	300	307	8	187	82.67
statlog	2000	2035	20	788	92.67
letter	4000	4176	20	3101	94.00
adult	4000	4055	50	3482	83.00
covtype*	3000	3210	200	1370	90.00

表 3　三种方法实验结果对比

数据集	方法一			方法二			本文方法		
	支持向量数	准确率/%	训练时间/s	支持向量数	准确率/%	训练时间/s	支持向量数	准确率/%	训练时间/s
wine	51.2	97.42	3.2	81	98.14	7.5	49.8	98.96	3.9
pima	198.9	76.82	33.3	418	76.35	171.1	197.9	79.48	36.3
statlog	842.5	90.51	1800.9	2452	92.07	16605.6	840.6	91.33	1991.1
letter	1703.0	82.81	3840.7	—	—	—	1383.0	82.87	5708.5
adult	3396.0	81.26	5123.6	—	—	—	3272.5	82.53	5214.6
covtype*	1198.0	85.04	5229.4	—	—	—	1407.0	86.91	8872.9

从表 3 中可以发现,本文方法除了在 statlog 数据集上的分类精度略差于方法二之外,在其他 5 个数据集上都得到了较好的分类精度。在支持向量个数方面,本文方法在 covtype* 数据集上的支持向量个数明显多于方法一,但在其他数据集上支持向量个数少于方法一和方法二。在训练时间上,方法一和本文方法都采用了采样策略,训练时间明显少于方法二。由于本文方法运用了分层聚类重采样的策略,因此在训练时间上要高于应用随机采样策略的方法一。总的来说,在训练样本数相同的情况下,分层聚类重采样算法的分类精度明显高于随机采样算法,其支持向量数目与训练时间均大大低于采用完整训练集训练的结果。通过表 2 与表 3 的对比可以发现,通过选取适当的采样数目和 k-means 划分块数,分层聚类重采样算法的分类准确率完全能达到完整训练集训练的结果,甚至更优。

3　结语

本文尝试采用约简训练集的方法来解决 SVM 对于大规模数据集的分类瓶颈问题。对训练集进行约简,主要依据两个条件:1)保证使用它训练时代价不高;2)保证使用它训练出的分类器有一定的分类精度。文中提出的基于分层聚类重采样约简策略在保证数据分布区域不变的情况下,对高密度区域数据进行约简,不会对支持向量的分布构成太大的影响。实验结果证实本文方法能够在降低学习代价的同时,很好地保证分类器的分类精度,其分类速度得到了较大的提高;另外,本文算法也在一定程度上抑制了过拟合的过学习现象。

参考文献:

[1] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社, 2004.

[2] 李红莲, 王春花, 袁保宗, 等. 针对大规模训练集的支持向量机的学习策略[J]. 计算机学报, 2004, 27(5): 715 – 719.

[3] CHEN P H, FAN R E, LIN C J. A study on SMO-type decomposition methods for support vector machines [J]. IEEE Transactions on Neural Networks, 2006, 17(4): 893 – 908.

[4] HUANG G B, MAO K Z, SIEW C K, *et al*. Fast modular network implementation for support vector machines [J]. IEEE Transactions on Neural networks, 2005, 16(6): 1651 – 1663.

[5] DONG J X, KRZYZAK A, SUEN C Y. Fast SVM training algorithm with decomposition on very large data sets [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(4): 603 – 618.

[6] CHEN G X, CHENG Y, XU J. Cluster reduction support vector machine for large-scale data set [C]// Proceedings of the 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. Piscataway: IEEE, 2008: 8 – 12.

[7] CERVANTES J, LI X, YU W. Support vector machine classification based on fuzzy clustering for large data sets [C]// MICAI'06: Proceedings of the 5th Mexican International Conference on Artificial Intelligence, LNCS 4293. Berlin: Springer, 2006: 572 – 582.

[8] LI D C, FANG Y H. An algorithm to cluster data for efficient classification of support vector machines [J]. Expert Systems with Applications, 2008, 34(3): 2013-2018.

[9] CERVANTES J, LI X, YU W, *et al*. Multi-class support vector machine for large data sets via minimum enclosing ball clustering [C]// Proceeding of the 4th International Conference on Electrical and Electronics Engineering. Piscataway: IEEE, 2007: 146 – 149.

[10] 陈光喜, 徐健, 成彦. 一种聚簇消减大规模数据的支持向量分类算法 [J]. 计算机科学, 2009, 36(3): 184 – 187.

[11] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [CP/OL]. [2012-10-10]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] FRANK A, ASUNCION A. UCI machine learning repository[EB/OL]. [2012-10-10]. http://archive.ics.uci.edu/ml.