DOI:10.16515/j.cnki.32-1722/n.2019.03.002

基于混合样本训练的并行层叠支持向量机研究

张洪胜,丁永红

(淮南联合大学计算机系,安徽 淮南 232038)

摘 要:层叠支持向量机将原始数据集随机划分为多个子集,对数据子集采取并行训练,可以有效提高分类器的训练效率。但其在将原始数据随机划分为多个训练子集时,可能会给各并行节点带来文本信息结构的不均衡,进而影响分类器的最终分类效果。提出了一种基于混合样本训练子分类器的训练模型,实验表明,基于混合样本训练的层叠支持向量机,可以较好地解决训练样本信息结构不均衡问题,保证层叠训练得到的分类器具有较好的精确度和稳定性。

关键词:文本分类;混合样本;层叠支持向量机

中图分类号:TP18;TP391 **文献标识码**:A **文章编号**:1672-755X(2019)03-0008-04

Research on Parallel Casecade SVM Based on Mixed Samples

ZHANG Hong-sheng, DING Yong-hong

(Huainan United University, Huainan 232038, China)

Abstract: Cascading support vector machine (SVM) divides the original data sets by randomly dividing them into multiple subsets. Parallel data subset training can effectively improve the training efficiency of the classifier. However, when the original data is randomly divided into multiple training subsets, it may bring the imbalance of various parallel node text information structure to each parallel node, and then affect the classification effect of the final classifier. In this paper, a training model based on mixed sample training subclassifier is proposed, and the experiment shows that the cascade support vector machine based on mixed sample training can solve the problem of unbalanced information structure of training samples, and ensure that the classifier obtained by cascade training has better accuracy and stability.

Key words: text classification; simulation sample; casecade support vector machine

支持向量机(Support Vector Machine,简称SVM)是20世纪90年代由V.Vapnic等在统计学习理论[1]的基础上提出的一种新的模式识别方法。该方法建立在结构风险最小化的基础之上,通过在学习系统的复杂性和学习能力之间寻找最佳折中点,以此寻求最好的推广能力。该方法在解决小样本、非线性以及高维度分类问题中表现出良好的性能,因而特别适合解决文本分类问题。但在海量文本数据处理方面,其存在着运算复杂度高、训练时间长、内存需求大等问题。随着云计算的兴起,MapReduce[2]等并行计算架构为支持向量机具备处理大规模文本数据能力提供了有力的技术支持。基于MapReduce的层叠支持向量机(Casecade SVM)作为并行SVM的一种,在大规模文本分类训练中体现出快速收敛的特点[3],但在对初始训练样本随机划分时,可能出现文本数据分块的信息结构不均衡,并由此导致全局支持向量减少,影响最后分类效果[4]。为此

收稿日期:2019-08-02

基金项目:安徽省教育厅自然科学重点项目(KJ2017A586)

作者简介:张洪胜(1968—),男,安徽淮南人,副教授,硕士,主要从事智能软件、并行计算、中文信息处理等研究。

文中提出一种用混合样本训练的层叠支持向量机文本分类模型，将初始训练样本随机划分到最上层的并行节点后，再在每个节点通过模拟样本随机生成算法随机生成一定规模模拟样本作为补充，形成文本信息结构较为均匀的混合训练样本，以提高层叠支持向量机文本分类的准确率和稳定性。

1　训练样本的获取

1.1　训练样本的向量表示

在机器学习与文本分类领域，一般采用向量空间模型[5]来表示文本信息，其基本思想是以词向量来表示文本：$(w_1, w_2, w_3, \cdots, w_n)$。其中，$w_i$为通过某种公式计算得到的第 i 个特征词的权重。特征词的选择是通过特征抽取算法得到，而权重的计算有相对词频和绝对词频两种方法[6]，绝对词频以词在文本中出现的次数表示，相对词频则利用 TF-IDF 等公式计算得到，TF-IDF 公式有多种，较为普遍的一种如式(1)所示：

$$W(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_i+0.01)]^2}} \tag{1}$$

其中，$W(t,d)$为词 t 在文本 d 中的权重，而 $tf(t,d)$为词 t 在文本 d 中的词频，N 为训练文本的总数，n_t为训练文本集中出现 t 的文本数，分母为归一化因子。

对于两类文本识别问题的支持向量机来说，用于训练的样本分为属于某一类别的正例样本和不属于这一类别的反例样本。无论是正例样本还是反例样本，在使用 SVM 算法训练之前，均须进行预处理，将其表示为向量的形式。

1.2　训练样本的模拟

1.2.1　*基于特征空间的模拟样本生成算法*　观察大量向量形式人工标注训练样本后发现，正例训练样本其权重值之和一般明显大于反例向量，而且权重值非 0 的特征在向量中出现的位置具有一定的随机性。根据这一现象，我们利用 TF-IDF 特征权重计算方法，设计了一种训练样本随机模拟生成算法，步骤如下：

第 1 步：准备一定数量的正例样本，对每个样本进行中英文分词并统计词频，根据词频大小，按照一定比例提取作为特征空间的特征，并做去重处理；

第 2 步：按照 TF-IDF 计算所有特征词的权重，为了得到特征词在正例所在类的权重，在计算时我们对公式(1)中 $tf(t,d)$的定义做了修改，将其由词 t 在文本 d 中的词频，改为 t 在所有正例样本中出现的词频；

第 3 步：随机生成样本中非 0 特征的个数 m(1～10)；

第 4 步：通过循环，随机生成 m 个特征在特征空间的位置 i，由此得到由 m 个特征组成的且特征出现位置随机的模拟样本；

第 5 步：根据需要生成样本的个数，重复第 3 步至第 4 步；

第 6 步：对于生成的每个向量形式的模拟样本，计算各维的权重之和 weightsum。如果满足条件“weightsum＜阀值”，则将该模拟样本作为反例样本，否则作为正例样本，由此得到用于训练学习的所有正反例样本。条件中的阀值可以通过实验选择一个使分类效果较好的值。

1.2.2　*模拟样本的训练效果*　表 1 的实验表明，利用上述模拟样本生成算法生成训练样本，训练得到的 SVM 分类器具有良好的分类性能。

表 1　模拟样本生成及 SVM 分类实验

反例生成条件	生成反例样本数	生成正例样本数	SVM 分类正确率/%
权重和＜0.5	58	942	85.4
权重和＜1.0	107	893	85.3
权重和＜1.5	233	767	84.0
权重和＜2.0	280	720	82.7

2 并行层叠支持向量机

2.1 并行 SVM 的 MapReduce 实现

MapReduce 是 Apache 软件基金会下大规模数据处理软件平台 hadoop 中的一个分布式计算框架，它采取“分而治之”的思想，把一个大的任务拆分成一系列小的任务并行处理，从而使任务快速得以解决。一个典型的 MapReduce 作业包括 Mapper 和 Reduce 两个过程，其中 Mapper 过程对原始数据进行拆分处理，Reduce 则对 Mapper 的结果进行汇总。基于 MapReduce 的并行 SVM，可以在 SVM 算法中实现并行处理，同时也可以通过对数据集的并行训练实现。例如，采用 PLATT 提出的序列最小优化(Sequential Minimal Optimization，SMO)算法[7]的 SVM，可以利用 MapReduce 在迭代前首先进行内积的并行运算[8]，以提高算法收敛的效率。数据训练集层面的并行 SVM，则首先在 Mapper 过程将原始训练数据集随机划分为多个子集，然后由 Reduce 过程对多个子集进行并行训练，最后将各子集的训练结果合并得到最终的支持向量。

2.2 层叠支持向量机训练模型

层叠支持向量机采用层级式的架构，将原始训练数据集随机划分为偶数个数据子集，并交由同一层不同计算节点并行训练，然后将同一层各节点的训练结果两两合并，进入下一层再次训练，直到最终得到整个训练样本集上的支持向量。以 4 层支持向量机为例，各层分组并发节点数分别为 8、4、2、1，其层叠分组训练模型如图 1 所示。

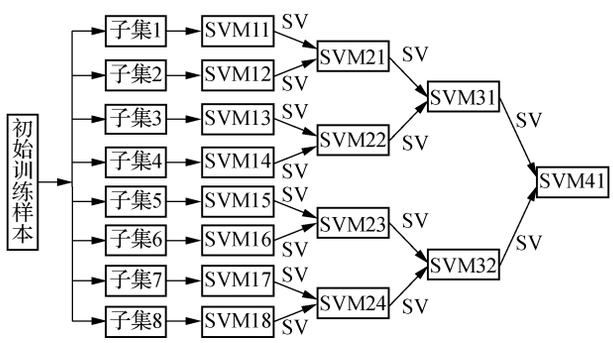

图 1 并行层叠支持向量机分组训练模型

在对原始数据集的随机划分中，只是保证了数据负载的均衡，但无法保证每个节点得到的正反例样本数量的均衡，而这种正反例样本数量上的不均衡体现了文本信息结构上的不均衡，可能导致子节点上出现全局支持向量丢失的现象，并影响最终模型支持向量的个数和模型的精确度，从而影响测试集上的预测准确率。

3 基于混合样本训练的层叠 SVM

层叠 SVM 在对初始训练样本随机划分时，可能使并行节点上的训练样本出现信息结构上的不均衡。对此，文本设计了一种基于混合样本训练的并行层叠 SVM 文本分类模型，该模型在原始训练数据随机划分到并行处理节点后，利用模拟样本生成算法，随机生成一定数量的模拟样本，作为对各并行节点训练子集的补充，随机模拟样本由于利用了统计规律设计算法生成，使所生成的样本具有很好的代表性和随机性，因而有助于提高各并行节点上训练样本在信息结构上的均衡性，从而提高层叠支持向量机文本分类的准确性和稳定性。基于混合样本训练的并行层叠 SVM 文本训练模型，以 4 层分组训练模型为例，如图 2 所示。

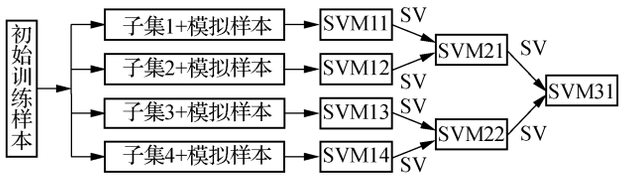

图 2 基于混合样本训练的并行层叠支持向量机

4　实验和结果

4.1　实验环境

实验使用 Dell R720 服务器搭建的 hadoop 伪分布式环境，双 E2520 CPU，64 G 内存，安装 XenServer 虚拟化服务器软件，在其中搭建由 1 个 namenode 节点和 6 个 datanode 节点组成的 hadoop 服务器集群。每个节点配置 4 G 内存、20 G 存储空间，安装 CentOS6.0 操作系统。

4.2　实验及结果分析

实验以 400 份人工标注训练样本为正例样本，通过特征抽取和 TF-IDF 权重计算生成特征空间，抽取特征 338 个。实验中的人工标注样本，是 NLP 文本分类语料库（复旦大学）计算机类的训练与测试集。

实验采取 4 层 SVM 分组结构，各层并发训练节点数量分别为 8、4、2、1，混合样本、非混合样本及纯模拟样本各进行 5 次实验，以观察这些样本层叠 SVM 各层文本分类的准确性和稳定性。实验结果如图 3 所示。

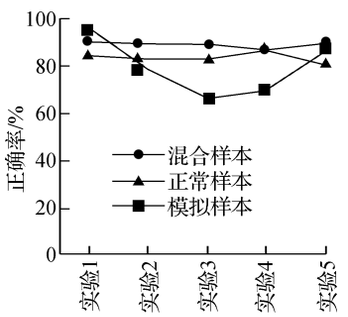

图 3　混合样本和非混合样本层叠 SVM 分类效果比较实验

正常样本实验的结果表明，在对训练数据进行随机初始划分后，如果不添加随机模拟样本，5 次训练得到的分类器，分类测试的准确率变动幅度较大。其主要原因是初始样本的随机划分，使各训练节点得到的训练样本结构信息出现了不均衡，因而子数据集在训练后可能导致部分全局支持向量的丢失，从而使层叠支持向量机在最后得到的支持向量并不是训练样本集上真正的最优解。

混合样本实验说明，在对初始训练样本进行随机划分后，在各数据节点利用模拟样本随机生成算法生成一定数量的模拟样本作为补充，可以有效避免初始划分带来的文本信息结构的不均衡问题，因而层叠 SVM 分类的准确率较高，同时稳定性也较好。

5　结　语

对于海量数据的文本分类问题，并行层叠 SVM 通过多层并行训练，可以实现训练过程的快速收敛，但其在初始训练数据随机划分时，可能导致各训练节点文本结构信息的不均衡，进而影响分类的效果。基于混合样本训练的层叠 SVM，通过在初始划分完成后，在各训练节点利用模拟样本生成算法，随机生成一定数量的模拟样本加以补充，利用混合样本进行训练，可以有效提高训练样本信息结构的均衡性，使并行层叠训练得到的分类器具有较好的分类效果。

参考文献：

[1] VAPNIK V. The nature of the statistical learning theory[M]. New York:Springer,1999

[2] HDFS. Architecture guide[EB/OL]. (2013－01－05)[2019－07－16]. http://hadoop. Apache. org/docs/hdfs/current/hdfs_design. html

[3] 张鹏翔，刘利民，马志强. 基于 MapReduce 的层叠分组并行 SVM 算法研究[J]. 计算机应用与软件，2015，32(3)：172－176

[4] 郭鹏辉. 层叠支持向量机优化及并行化实现[D]. 兰州：兰州大学，2018

[5] 张东礼，汪东升，郑伟民. 基于 VSM 的中文文本分类系统的设计与实现[J]. 清华大学学报(自然科学版)，2003，43(9)：1288－1291

[6] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001(9)：23－26

[7] PLAT J C. Fast training of support vector machines using sequential minimal optimization[M]. Cambridge M. A. :MIT Press,1999:185－208

[8] 杨鹤标，黄文青，陈锦富. 基于 MapReduce 的 SVM 改进算法及在邮件过滤中的实现[J]. 无线通讯技术，2013(2)：52－56

(责任编辑：湛　江)