

一种处理不平衡大数据的并行随机森林算法

秦 静, 钱雪忠, 王卫涛, 谢国伟, 宋 威

(江南大学 物联网技术应用教育部工程研究中心, 江苏 无锡 214122)

摘 要: 基于 MapReduce 机制的并行随机森林算法 MR_RF 在处理不平衡大数据时, 由于数据本身正类密度低且算法又以全局最优标准选择分割点, 而导致正类有误分为负类的趋势, 降低了分类效率. 本文提出了一种改进的并行随机森林算法(SBWMR_RF), 该算法利用分层自助抽样方法增大对少数类的抽样数量, 同时考虑正负类不同的误分代价, 动态计算每个分区的代价敏感矩阵, 将其引入到构建基分类器的关键步骤, 弱化数据偏置的影响. 实验证明 SBWMR_RF 算法提高了对不平衡大数据的分类准确率, 没有出现过拟合现象, 在极不平衡环境下优势明显.

关键词: 不平衡大数据; MapReduce; 随机森林; 代价敏感; 分层自助抽样

中图分类号: TP311.1

文献标识码: A

文章编号: 1000-7180(2017)04-0022-06

A Algorithm for Unbalanced Big Sata Using Paralleled Random Forest

QIN Jing, QIAN Xue-zhong, WANG Wei-tao, XIE Guo-wei, SONG Wei

(Engineering Research Center of Internet of Things Technology Applications,
Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: Paralleled random forest algorithm based on MapReduce(MR_RF) which constructing trees on partitions is a classic ensemble algorithm for big data classification. However, when encountering imbalanced big data, it's performance will decrease with the tendency of positive samples misclassified because of the low density of positive samples themselves and the algorithm's global optimal criteria for choosing split points. In this paper, An improved paralleled random forest called SBWMR_RF is proposed. It adopts stratified bootstrap to increase the minority during sampling. At the same time, cost-sensitive thought is applied through the key steps of tree construction, modify the distribution of the minority. The experiments prove that SBWMR_RF can effectively classify unbalanced big data especially in extremely unbalanced data scenario without overfitting but high speedup.

Key words: unbalanced big data; MapReduce; random forest; cost-sensitive; stratified bootstrap

1 引言

在大数据时代不平衡学习面临着更为严峻的挑战, 如何快速高效地对海量不平衡数据进行分析处理, 成为数据挖掘领域的研究热点^[1].

处理大数据的技术重点是实现快速、可伸缩的并行化分析处理. 现在主流的手段是遵循 MapReduce 流程, 先将大数据分成多个小的子集, 然后对子集处理的结果进行合并^[2].

随机森林算法(RF)^[3]是 Leo Breiman 在 2001 年提出的结合自助抽样(bootstrap)和子空间思想的集成分类算法. 而基于 MapReduce 的并行随机森林算法(MR_RF)^[4]采用分治策略, 将基分类器建立在每个分片(partition)的抽样数据上, 更加强化了数据偏置的影响^[5]. 另一方面 RF 在构建树时采用全局最小 gini 值作为选择分裂点的标准, 易忽略正类, 有将正例分为负例的趋势^[6]. 集成分类处理不平衡数据时从数据层面出发改进, 主要是获得大致平

衡数据子集如 OverBagging、UnderBagging、easy-Ensemble 等^[7],也有从改进算法层面出发,重构训练集的同时,引入代价敏感因子,设计代价敏感基分类器^[8].

有鉴于此,本文提出一种综合的方法,SBWMR_RF(Stratified bootstrap and weighed random forest on mapreduce)算法. 其将分层自助抽样方法和代价敏感思想引入到 MR_RF 中,结合 MapReduce,从调整数据分布和调节算法两个方面对不平衡大数据分而治之. 首先对每个分区数据集的正例和负例分别进行自助抽样,保证随机性的同时增加正类的信息量. 其次计算每个分区的代价敏感矩阵,将敏感代价用于基分类器构建的核心步骤,如计算最优分割点,叶子节点所属类别,预测时加权投票等,构建代价敏感的决策树. 最后对 SBWMR_RF 模型利用 F_β value 和 G-mean 指标检验分类效果. 实验证明本文提出的 SBWMR_RF 可以有效处理地不平衡大数据分类问题.

2 基于 MapReduce 的并行随机森林

MR_RF 是由 Mahout 开源项目实现的基于 MapReduce 机制的并行化随机森林算法. 其主要内容有:

(1)MapReduce 机制:能够将工作流分配到大规模集群节点之上,然后各个节点并行计算. 作为一种线性可伸缩的编程模型,将任务分为 2 个处理阶段,Map(映射)阶段和 Reduce(约简)阶段,每个阶段都以 $\langle \text{key}, \text{value} \rangle$ 键值对作为输入输出.

(2)RF 算法对数据集 $D_{[N,S]}$ k 次自助抽样得到子集 $\{D_1, D_2, \dots, D_k\}$; 在每一个 D_i 上建立一棵 CART 树,过程中随机抽取 p 个属性 ($p \ll S$),取当前最小 Gini 值作为分裂点,直到满足停止条件,如所有数据都是一类、未处理数据少于一定的量等就停止,得到森林 $\{h_1, h_2, \dots, h_k\}$,预测时采用简单投票机制.

(3)MR_RF 步骤:算法分为模型构建和数据分类预测两个阶段. 森林构建:基于分治思想,将数据分区(一般按 64M/分区)传递给不同的数据节点做计算,在 map 函数中,针对分区数据的 bagging 子集用 RF 算法构建基分类器,map 任务并行运行,多个决策树汇聚成森林. 分类预测:首先将测试集分成几个互相独立的子集分发到不同的数据节点,在 map 任务中,遍历森林里的每一棵树,采用简单投票,取预测结果的众数作为最终的分类结果. 将结果

汇聚,计算分类准确率.

3 SBWMR_RF 算法

SBWMR_RF 算法是建立在 hadoop 平台上的改进的并行随机森林算法. 一方面将分层自助抽样替换 RF 的自助抽样,增加正类的抽样数量概率,进而调节算法对数据的偏置,另一方面在每个分区树构建的过程中加入敏感代价,这能改善由于全局搜索策略而导致正类被误分的问题,最终使得在大数据环境下的不平衡分类结果更加准确与高效.

3.1 分层自助抽样

分层抽样技术是大规模数据管理常用的技术^[9]. 核心思想是在不平衡数据集的二元分类中,对于正类,负类两层数据分别自助抽样,增大对正类的抽样概率. 假设我们有数据集 $D_{[N \times S]}$,其中负类有 H 个, $N-H$ 个是正类 ($N > H \gg N-H$),随机抽取一个数是负类的概率是 $\frac{H}{N}$ 则自助抽样后数据集中全是负类的概率是

$$\left(\frac{H}{N}\right)^n = \left(1 - \frac{N-H}{H}\right)^n$$

在极不平衡的数据环境下,即 $N > H \gg N-H$ 时有很大的可能性 bagging 后的数据全是来自负类,所以利用分层自助抽样,对正类和负类每一层都进行自助抽样,从而确保有一定的正类出现在抽样集中.

3.2 代价敏感方法

除了增加对正类的抽样概率,还需修改算法在数据集上的偏置,改善由于全局搜索策略而导致正类被误分的问题. 所以引入代价敏感学习方法. (1) 重构训练集. 不改变已有算法,但根据样本的不同错分代价给每一个样本赋一个权值,接着按不同权重对原始样本集进行重新统计. (2) 引入代价敏感因子设计出代价敏感的决策树算法. 通常对小样本赋予较高的代价,大样本赋予较小的代价,期望以此来平衡样本之间的数目差异.

其关键步骤是获取代价敏感矩阵. 记正类被误分代价为 $C(+, -)$, 相对的负类被误分代价为 $C(-, +)$ 且 $C(+, -) > C(-, +)$, $C(+, +) = C(-, -) = 0$. 动态计算每一个分区的代价敏感矩阵设置

$$C(+, -)_k = \frac{H_k}{N_k - H_k}, C(-, +) = 1 \quad (1)$$

在统计时正类和负类的比率大致平衡即

$$\frac{(N-H) \times \frac{H}{N-H}}{H \times 1} = 1:1$$

将每个类别的不同误分代价引入到 CART 树的构建中(1) 在构建计算树的计算最佳分裂点时加入代价敏感信息,如式(2);(2) 计算叶子节点所属哪个类别,如式(3);(3) 计算叶子节点的权值,如式(4),用于分类预测阶段的有权投票。

$$\text{Gini}(D, A) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad (2)$$

表示在分割点 A 的条件下,集合 D 的不确定性。 D_1 和 D_2 是按照是否是 A 值将集合 D 分成的两部分。其中 $\text{Gini}(D) = 1 - \sum_{k=1}^K p_k^2$, $p_k =$

$$\frac{|D_k| \times c_k}{\sum_{i=1}^K (|D_i| \times c_i)}, D_k \text{ 是 } D \text{ 中属于第 } k \text{ 类的样本子集}, c_k \text{ 是第 } k \text{ 类的误分代价}, K \text{ 是类的个数}.$$

$$I = \max\{|D_1| \times c_1, |D_2| \times c_2, \dots, |D_k| \times c_k\} \quad (3)$$

表示叶子节点所属类别。

$$W = \frac{\sum_{k=1}^K |D_k| \times c_k}{\sum_{i=1}^K |D_i|} \quad (4)$$

表示每个叶子节点的权重,是实例的代价和与实例个数总和的比值。

3.3 SBWMR_RF 模型构建流程

分类模型阶段由 Initial、Map 和 Final 三个步骤组成。

Initial 阶段:首先在 Initial 时数据分区到不同的数据节点上,然后计算每个分区的代价矩阵,如公式(1), $C(+, -)$ 采用动态的局部不平衡指数。

Map 阶段:输入键值对 $\langle \text{key}, \text{value} \rangle$, key 为编码一个实例的二进制流, value 是每个实例的具体数据。对各个分区数据进行分层自助抽样,然后组成最后的抽样数据集, $\{d_1, d_2, \dots, d_g\} (g < k)$ 构建代价敏感 CART 树。输出键值对 $\langle \text{key}', \text{value}' \rangle$, key' 由树 id 和分区 id 组成, value' 是一个基分类器。

Final 阶段:汇总各棵树形成 SBWMR_RF 随机森林 $\{wh_1, wh_2, \dots, wh_k\}$ 。图 1 是 SBWMR_RF 构建树的流程图。

3.4 SBWMR_RF 分类预测流程

分类预测阶段也是由 Initial、Map 和 Final 组成。

Initial 阶段:将测试集分成几个互相独立的子

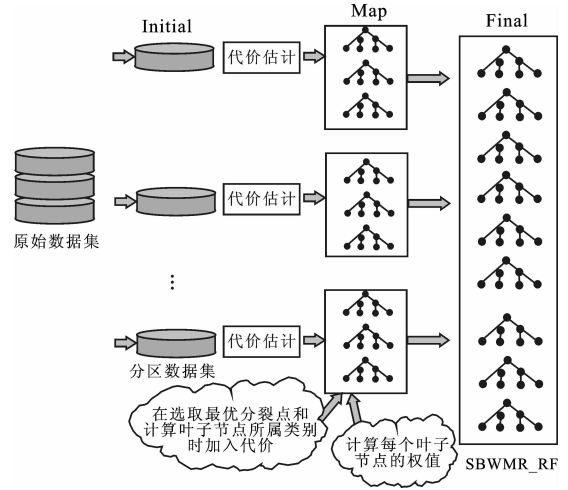


图 1 SBWMR_RF 构建树的流程图

集并分发到不同的数据节点,

Map 阶段:是主要的预测方法,输入键值对 $\langle \text{key}, \text{value} \rangle$, key 为编码实例的二进制流, value 是实例的具体数据。遍历所有树,采用加权投票的方式,如公式(4),取得最终的预测结果。输出键值对 $\langle \text{key}', \text{value}' \rangle$, key' 是实例的实际类别 value' 是一个实例的预测类别。

final 阶段:将结果汇聚,计算不同的指标值。图 2 是 SBWMR_RF 分类流程图。

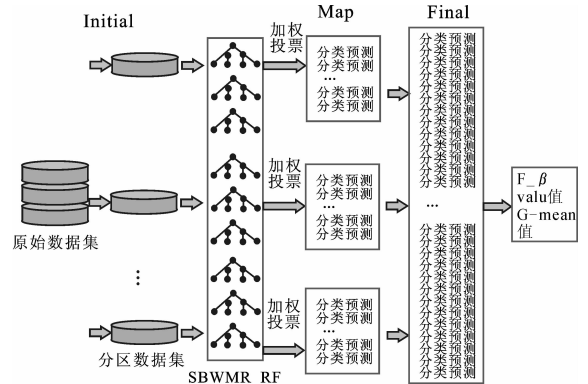


图 2 SBWMR_RF 分类流程图

4 实验验证

4.1 实验环境

本文搭建的 Hadoop 平台为 4 台集群节点,由 1 明 000 书 M 宽带互联。配置均为 CPU:Deon L5639 的、内存:4 GB、缓存:12 MB、硬盘:200 GB、Hadoop 版本:Hadoop 2.0.0 CDH 4.5、OS 版本:Centos 7.0、Mahout 版本:Mahout 0.7 CDH 4.5。

4.2 实验数据集

本文选取的数据集为 KDD Cup 1999 和 the

Record Linkage Comparison Patterns(RLCP).

同时本文,将 KDD Cup 1999 这个多分类数据集按照不同类别形成不平衡二元分类大数据集,从上至下数据量依次增大,而 RLCP 本身是二分类数据集.数据集的具体信息见表 1.

表 1 不平衡数据集				
Datasets	NumInstance	NAttr	Class($n:p$)	IR
kddnormal_VS_U2R	972833	41	972781:52	18707.327
kddnormal_VS_R2L	973907	41	972781:1126	863.926
kddnormal_VS_PRB	1013883	41	972781:41102	23.667
kddDOS_VS_U2R	3883422	41	3883370:52	74680.192
kddDOS_VS_R2L	3884496	41	3883370:1126	3448.819
kddDOS_VS_PRB	3924472	41	3883370:41102	94.481
kddDOS_VS_normal	4856151	41	3883370:972781	3.992
RLCP	5749132	11	5728201:20931	273.671

其中 NumInstance 表示实例的数量、NAttr 表示属性的数量、Class($n:p$)表示负类与正类的数量、IR 是负类与正类数量的比值,其表示全局不平衡指数.

4.3 不平衡数据评价指标

传统的分类学习方法中,一般采用分类精度作为评价指标.然而该指标对于不平衡数据集不适用,因为正类比例如果不足 1%时即使正类被分为负类,精度依旧可以达到 99%,但这样的分类器是没有实际意义的.为此本文采用常用的 F_β value 和 G-mean 标准作为评价指标.这两个指标都是由混淆矩阵而来,具体见表 2. 其中,TP 表示正类样本判为正类的数目、TN 表示负类样本判为负类的数目、FN 与 FP 分别表示判断错误实际为正类和负类的样本.

表 2 混淆矩阵		
分类	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

F_β value 指标是综合考虑查全率和查准率的分类评价指标,(+)表示越大越好:

$$F_\beta\text{value}(+) = \frac{(1 + \beta^2) * \text{recall} * \text{precision}}{\beta^2 * \text{recall} + \text{precision}}$$

(5)

式中,precision 为查准率,recall 为查全率, β 取值为

$[0,\infty]$. F_β value 可以用来考察召回率和精度之间的折中. $\beta<1$ 时强调 precision 的作用; $\beta>1$ 时强调 recall 的作用. precision 和 recall 的定义如下:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

G-mean 值表示的是正类分类精度和负类分类精度的几何平均值. (+) 表示越大越好. 其定义如下:

$$G\text{-mean}(+) = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FP}} * \frac{\text{TN}}{\text{FP} + \text{TN}}}$$

(6)

G-mean 是保持负类,正类分类精度平衡的情况下最大化两类的精度,即只有在负类和正类的分类精度同时都高的情况下,G-mean 的值才最大. 本文将使用 G-mean 来衡量数据集整体的分类性能.

4.4 实验设计

SBWMR_RF 的参数集中在森林构建阶段. 其中,树的最大深度默认为无限, m 表示每次选取的属性数量为 $\text{sqrt}(M)$, M 为属性总数, K 表示数的数量设置为 100. F_β value 中 β 取全局 IR 值^[10]. 本文采用 5 折分层交叉验证. 所有的数据结果是五次结果的平均值.

4.5 实验结果分析

(1)分类准确率分析

首先比较 SBWMR_RF 在不同数据集上的准确率. 表 3 是两种算法在不同数据集上的 G-mean 值. 从表中可以看出以下结论.

表 3 SBWMR_RF 和 MR_RF 在 8 个数据集上 G-mean 值对比表				
Datasets	MR_RF		SBWMR_RF	
	GMtr	GMts	GMtr	GMts
Kddnormal_VS_U2R	0.556	0.553	0.973	0.952
Kddnormal_VS_R2L	0.933	0.923	0.977	0.977
Kddnormal_VS_PRB	0.993	0.991	0.996	0.992
kddDOS_VSU2R	0.556	0.548	0.886	0.836
kddDOS_VS_R2L	0.991	0.991	0.978	0.975
kddDOS_VS_PRB	1.000	1.000	0.997	0.996
kddDOS_VS_normal	1.000	1.000	1.000	1.000
RLCP	0.116	0.116	0.933	0.932
Average	0.735	0.732	0.968	0.958

第一:绝大多数数据集的训练数据集和测试数

据集的 G -mean 值都没有很大差距,说明 SBWMR_RF 没有出现过拟合. 第二:分析不平衡程度极高的数据集如 kddDOS_VS_U2R,Kddnormal_VS_U2R 两个数据集. SBWMR_RF 准确率在较 MR_RF 有明显提高. SBWMR_RF 在极不平衡数据集下优势明显. 第三:分析不平衡程度相对低的数据集,MR_RF 和 SBWMR_RF 的性能都很好,但 SBWMR_RF 提高了正类识别率的同时也增加了负类误分为正类的概率,导致 G -mean 值较 MR_RF 有些许降低.

综合考虑 8 个数据集,SBWMR_RF 分类性能稳定,效率较高,由于分层自助抽样和加权建树使得它在极不平衡的数据集下分类优势明显. 表 4 是两种算法在不同数据集上的 F_{β} value 值. 从表 4 可以看出以上类似结论. 说明不同指标验证下,SBWMR_RF 的综合分类效率都比 MR_RF 有一定的提升.

表 4 SBWMR_RF 和 MR_RF 在 8 个数据集上

Datasets	MR_RF		SBWMR_RF	
	F_{β} tr	F_{β} ts	F_{β} tr	F_{β} ts
Kddnormal_VS_U2R	0.473	0.447	0.952	0.952
Kddnormal_VS_R2L	0.869	0.853	0.973	0.975
Kddnormal_VS_PRB	0.993	0.991	0.996	0.992
kddDOS_VSU2R	0.473	0.462	0.786	0.700
kddDOS_VS_R2L	0.987	0.986	0.957	0.951
kddDOS_VS_PRB	1.000	0.999	0.978	0.976
kddDOS_VS_normal	1.000	1.000	0.999	0.999
RLCP	0.014	0.014	0.992	0.991
Average	0.726	0.719	0.954	0.942

(2)算法的时间性能比较

根据每个数据集五次模型构建时所用时间的平均值,得到了图 3 的时间性能对比图. 其中,数据集中的数据量从上至下依次越来越大. 从图 3 的实验结果,可以看出 SBWMR_RF 时间性能并不占优势,这是由于数据分层后抽样增加了额外的 I/O 操作,同时构建树时加入代价敏感信息,增加了计算时间.

然而,随着数据量的不断增大,可以看出 SBWMR_RF 算法的并行处理优势更加明显. 当数据分片数是 2 的时候,由于数据集相对较小,此时集群资源利用率不高,时间主要耗费在启动 map 上和算

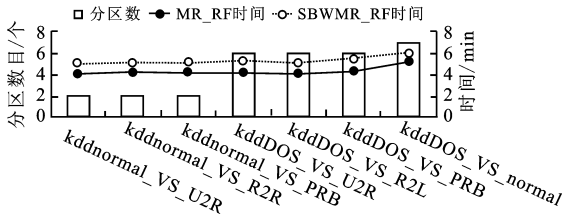


图 3 时间性能对比图

法计算;当分区数增大到 6 时,SBWMR_RF 运行时间并没有明显增加,这是因为集群的工作节点被均衡调度,多个任务并行运行,所以最终的时间取最慢计算时间;当分区数为 7 时,计算时间只是在初始化期间有所增多,而最终的总体时间基本一致. 综上分析,可以看出 SBWMR_RF 算法可以适应大数据对不平衡分类问题进行处理,同时能在更加充分地利用集群资源的基础上,更优异地处理大数据集.

(3)SBWMR_RF 算法加速比实验

图 4 给出了 SBWMR_RF 算法与对比算法的加速比实验结果. 加速比可以衡量算法的并行性能,它指的是同一个任务单节点运行时间与多节点运行时间的比率. 从结果可以看出在 3 个数据集上,本文提出的 SBWMR_RF 算法的加速效果均较好,这表明 SBWMR_RF 算法具有良好的并行性可以适应大数据处理要求.

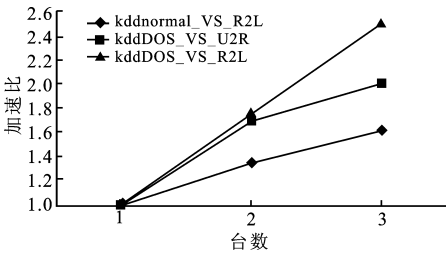


图 4 三个数据集上 SBWMR_RF 算法的加速比

5 结束语

本文提出了基于 MapReduce 机制的改进的并行化随机森林 SBWMR_RF 算法. 该算法分别利用分层自助抽样和代价敏感方法改善数据偏置和算法偏置,最终改善了 MR_RF 算法在处理不平衡大数据过程中由于密度低而导致分类效率降低的问题. 实验结果表明,对不平衡大数据分类效率有所提高. 但是,给定的代价敏感计算公式需要进一步调优,使得正类和负类都得到好的识别;另一方面希望在更新的大数据平台上如 Spark 上开展研究,进一步提升分类速度.

参考文献:

- [1] López V, Fernández A, García S, et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics[J]. Information Sciences, 2013, 250(11): 113-141.
- [2] 李建江, 崔健, 王聘, 等. MapReduce 并行编程模型研究综述[J]. 电子学报, 2011, 39(11):2635-2642.
- [3] Breiman L. Random forest[J]. Machine Learning, 2001, 45(1):5-32.
- [4] Apache Mahout. Classifying with random forests[EB/OL]. [2016-08-26]. <http://mahout.apache.org/users/classification/partial-implementation.html>, 2016. 6. 14/2017. 7.
- [5] Genuer R, Poggi J M, Tuleau Malot C, et al. Random forests for big data[J]. arXiv preprint arXiv:1511.08327, 2015.
- [6] Chen C, Breiman L. Using random forest to learn imbalanced data[J]. Telkomnika Indonesian Journal of Electrical Engineering, 2004, 45(1):5-32.
- [7] Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models[C]// IEEE Symposium on Computational Intelligence and Data Mining. Nashville, TN, USA, 2009:535-548.
- [8] 谷琼, 袁磊, 熊启军, 等. 基于非均衡数据集的代价敏感学习算法比较研究[J]. 微电子学与计算机, 2011, 28(8):146-149.
- [9] Kamiński C, Kelner J, Mariz D, et al. A stratified traffic sampling methodology for seeing the big picture[J]. Computer Networks the International Journal of Computer & Telecommunications Networking, 2008, 52(14):2677-2689.
- [10] Chawla N V, Cieslak D A, Hall L O, et al. Automatically countering imbalance and its empirical relationship to cost[J]. Data Mining & Knowledge Discovery, 2008, 17(2):225-252.

作者简介:

秦 静 女, (1992-), 硕士研究生. 研究方向为数据挖掘. E-mail: emily139617@126.com.

钱雪忠 男, (1967-), 副教授. 研究方向为数据库、数据挖掘.

王卫涛 男, (1989-), 硕士研究生. 研究方向为数据挖掘.

谢国伟 男, (1991-), 硕士研究生. 研究方向为计算机应用技术、数据挖掘.

宋 威 男, (1981-), 博士, 副教授. 研究方向为数据挖掘、人工智能和模式识别、信息检索.

(上接第 21 页)

- [4] Li M, Lian X C, Kwok J T, et al. Time and space efficient spectral clustering via column sampling [J]. IEEE Computer Vision and Pattern Recognition, 2011, 32(14):2297-2304.
- [5] 张顺龙, 库涛, 周浩. 针对多聚类中心大数据集的加速 K-means 聚类算法[J]. 计算机应用研究, 2016, 33(2):413-416.
- [6] Fraing G, Sohler C. A fast K-means implementation using corsets[J]. International journal of computational Geometry & Applications, 2008, 18(6):605-625.
- [7] Romero E. Using the leader algorithm with support vector machines for large data sets[J]. Lecture Notes in Computer Science, 2011(6791):225-232.
- [8] Viswanath P, Babu VS, Rough-DBSCAN: a fast hybrid density based clustering method for large data set [J]. Pattern Recognition Letters, 2009, 30(16):1477-1488.

作者简介:

刘 岩 女, (1976-), 硕士, 工程师. 研究方向为网络管理及计算机应用. E-mail: l_y_2016@sina.com.

王存睿 男, (1980-), 硕士, 讲师. 研究方向为软件与理论.