一种多标签随机均衡采样算法*

李思豪, 陈福才, 黄瑞阳
(国家数字交换系统工程技术研究中心, 郑州 450002)

摘 要: 为解决多标签学习中数据不平衡、传统重采样过程标签样本集相互影响以及弱势类信息大量重复和强势类信息大量丢失的问题,提出多标签随机均衡采样算法。该算法在多标签的条件下提出随机均衡采样思想,充分利用强势类和弱势类信息来平衡数据冗余和损失;优化样本复制和删除策略,保证不同标签重采样过程的独立性;提出平均样本数,保持数据的原始分布。实验在三个数据集下对比了三种多标签重采样算法的性能,结果表明,0.2 和 0.25 是所提算法的最佳重采样率,且该算法尤其适用于不平衡度较高的数据集,与其他方法相比具有最好的性能。

关键词: 多标签学习; 数据不平衡; 平均样本数; 随机均衡采样
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2017)10-2929-04
doi:10.3969/j.issn.1001-3695.2017.10.011

Multi-label random balanced resampling algorithm

Li Sihao, Chen Fucai, Huang Ruiyang
(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: To deal with the class imbalance in multi-label learning, the interaction between different label sets, the information redundancy of minority classes as well as the loss of majority classes that existed in traditional multi-label resampling, this paper put forward a multi-label random balanced resampling algorithm. The algorithm proposed the random balanced resampling method to make use of minority and majority information to balance the redundancy and the loss, improved replication and deletion strategy to ensure the independence of the resampling process of different labels, and maintained the original distribution of dataset with newly proposed mean instance size. Experiment results show that the proposed method is especially suit for datasets with higher imbalance ratio and achieves the best performance, 0.2 and 0.25 are the best resampling ratios in it.

Key words: multi-label learning; class imbalance; mean instance size; random balanced resampling

数据不平衡是广泛存在于数据集中的异类样本数相差较大的现象。其中,样本数较多的类称为强势类,对应强势标签,样本数较少的类称为弱势类,对应弱势标签。传统的多标签学习(multi-label learning,MLL)大多建立在数据集中类分布基本平衡的假设上,数据的不平衡导致分类边界偏向于弱势类,使得样本更容易获得强势标签[1]。然而在实际应用中,错分弱势类样本虽然不会对系统总的准确率产生太大影响,却会带来更大的代价[2,3]。所以,数据的不平衡问题是 MLL 的主要挑战之一。当前主要从两个层面解决多标签数据的不平衡问题:在算法层面,是对已有的多标签学习算法进行改进来改善弱势类样本的识别准确率[4~6];在数据层面,是利用重采样方法,通过人工构造弱势类样本来增加弱势类样本数(过采样,over-sampling)或减少强势类样本数(欠采样,under-sampling),使标签样本数达到平衡[7~9]。随机重采样(random re-sampling,RRS)独立于机器学习算法,是最简单的重采样技术,它随机选择弱势类样本进行复制或随机选择强势类样本进行删除。Charte 等人[10]将 MLL 的 LP 算法与传统的随机重采样结合,提出 LP-RUS 和 LP-ROS 算法,首次解决了传统的随机重采样算法无法应用于多标签数据集的问题。鉴于 LP-RUS 和 LP-ROS 较高的算法复杂度,他们又提出了多标签随机过采样(ML-ROS)和多标签随机欠采样(ML-RUS)算法[11](以下简称多标签随机重采样(ML-RRS))。ML-RRS 考虑了数据集中每个标签的分布特性,在提高了分类性能、降低算法复杂度的同时,单独进行过采样和欠采样,导致存在标签样本集相互影响以及弱势类信息大量冗余或强势类信息大量丢失的问题。基于此,本文在优化 ML-RRS 算法采样策略,保证不同标签采样过程相互独立的基础上,同时进行过采样和欠采样,提出多标签随机均衡采样技术(multi-label random resampling,ML-RBS),使数据以更均衡的方法达到平衡。

1 相关工作

1.1 多标签数据集不平衡程度衡量指标

定义 1 设 $\chi = X_1 \times X_2 \times \cdots \times X_d$ 表示 d 维特征空间,m 为样本数,q 为标签数,定义多标签数据集 $D = \{(x_i, Y_i) \mid 0 \leqslant i \leqslant m, Y_i \subseteq Y\}$。其中,$x_i = \{x_{i1}, x_{i2}, \cdots, x_{id}\}$ 表示一个样本,x_{il} 表示样本 x_i 的第 l 维特征值,$Y = \{y_1, y_2, \cdots, y_q\}$ 表示 D 的标签集,Y_i 表示样本 x_i 所属的标签集合。

收稿日期: 2016-07-05; **修回日期**: 2016-09-01 **基金项目**: 国家自然科学基金资助项目(61171108);国家“973”计划资助项目(2012CB315901,2012CB315905);国家科技支撑计划资助项目(2014BAH30B01)

作者简介:李思豪(1991-),男,云南昆明人,硕士,主要研究方向为网络大数据分析技术、多标签学习(1242100831@qq.com);陈福才(1974-),男,研究员,硕导,主要研究方向为电信网关防、网络大数据分析技术;黄瑞阳(1986-),男,助理研究员,博士,主要研究方向为网络大数据分析技术、大数据分布式处理技术.

对于二元数据集,常用强势类样本数与弱势类样本数之比作为不平衡度(imbalance ratio,IR),来衡量数据的不平衡程度[12]。多标签分类问题中,强势类或弱势类包含多个标签,也即数据集中一个样本可能既属于强势类也属于弱势类,不平衡度的衡量需要同时考虑多个标签。基于此,文献[10,11]提出了单标签不平衡度(imbalance ratio per label,IRL)和平均不平衡度(mean imbalance ratio,meanIR),作为多标签数据集不平衡程度的衡量指标。

IRL 表示为最大(本文用标签对应的样本数来衡量标签的大小,标签对应的样本数越多,标签越大,反之越小)。强势标签的样本数与当前标签的样本数之比,如式(1)所示。$IRL(y) \geqslant 1$,当且仅当 y 为最大强势标签时取等号。IRL 考察了单个标签的分布,数值越大,当前标签的不平衡度越大。

$$IRL(y) = \frac{\operatorname{argmax}_{y'=Y_1}^{Y_q}\left(\sum_{i=1}^{m} h(y', Y_i)\right)}{\sum_{i=1}^{m} h(y, Y_i)} \tag{1}$$

其中:$h(y, Y_i) = \begin{cases} 1 & y \in Y_i \\ 0 & y \notin Y_i \end{cases}$。

meanIR 表示数据集中各标签的平均不平衡程度,用多标签数据集中各标签的 IRL 的均值来衡量,如式(2)所示。meanIR 考察了数据集不平衡程度的平均水平。

$$meanIR = \frac{\sum_{y=Y_1}^{Y_q} IRL(y)}{q} \tag{2}$$

1.2 多标签随机重采样算法

定义 2 多标签数据集中,对于标签 $y \in Y$,若 $IRL(y) > meanIR$,则 y 属于弱势标签,对应样本属于弱势类样本;若 $IRL(y) \leqslant meanIR$,则 y 属于强势标签,对应样本属于强势类样本。

ML-ROS 算法伪代码如算法 1 所示。

算法 1 多标签随机过采样算法 ML-ROS

输入:原始不平衡多标签数据集 mlData,过采样率 P。

输出:过采样后的多标签数据集 returnedData。

1 计算待复制样本数 samplesToClone = |mlData| × P,得到数据集的标签集合 L;

2 由式(2)计算原始数据集 mlData 平均不平衡度 meanIR;

3 将 L 中所有弱势标签及其样本集归入弱势类数据集 minBag 中;

4 执行步骤 5 直至 samplesToClone = 0;

5 对于 minBag 中的每一个标签 y':随机选择该标签的一个样本进行克隆,此时如果 $IRL(y') \leqslant meanIR$,就将 y' 及其样本集移出 minBag,同时 samplesToClone - -;

6 返回 returnedData 进行进一步操作;

ML-RUS 和 ML-ROS 算法的不同是在步骤 5 中,ML-RUS 算法随机选择强势类标签的一个样本进行删除,直至该标签的单标签不平衡度和 meanIR 相等。

由算法 1 可以看出,ML-RRS 算法虽然很好地利用了原始多标签数据集的信息,但是:a)由于一个样本可能同时包含弱势标签和强势标签,对一个标签样本复制或删除的同时,也影响到其他标签的大小,从而对原始数据分布造成破坏;b)由于依序平衡弱势标签,完成采样时靠后的标签可能尚未达到平衡状态;c)单独进行过采样和欠采样易造成大量的弱势类信息冗余和强势类信息丢失,两者对分类性能产生不利影响[12]。

2 多标签随机均衡采样技术

2.1 平均样本数

定义 3 平均样本数。满足 $IRL(y_m) = meanIR$ 的标签 y_m ($y_m \in Y$)的样本数。

性质 1 多标签数据集中,平均样本数唯一,且可由任意标签的样本分布求出。

证明 设多标签数据集中最大标签所对应样本数为 max,y_m 为具有平均样本数的标签。由式(2)及定义 2,得式(3)。由式(1),对任意标签 $y \in Y$,得式(4),将式(4)和(3)作比可得式(5)。由于 max 在多标签数据集中唯一,所以性质 1 得证。

$$meanIR = IRL(y_m) = \frac{max}{meanInstance} \tag{3}$$

$$IRL(y) = \frac{max}{|instancesof(y)|} \tag{4}$$

$$meanInstance = \frac{max}{meanIR} = \frac{IRL(y)}{meanIR} \times |instancesof(y)| \tag{5}$$

由定义 2 和式(5)易得性质 2。

性质 2 多标签数据集中,对任意标签 $y \in Y$,如果 y 是弱势标签,则 $|instancesof(y)| < meanInstance$,如果 y 是强势标签,则 $|instancesof(y)| > meanInstance$。

2.2 算法描述

ML-RBS 算法流程如图 1 所示。

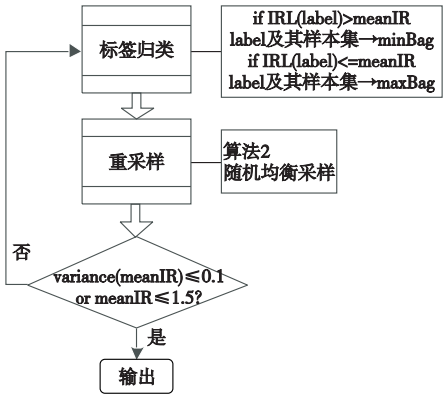


图1 ML-RBS算法流程

a)根据定义 2 划分弱势标签集(minBag)和强势标签集(majBag)。

b)利用算法 2 所示的 RBS 算法进行数据重采样。该算法随机选择一个标签,如果是弱势标签且样本数小于平均样本数,就进行过采样;如果是强势标签且样本数大于平均样本数,就进行欠采样,直至待重采样样本数为 0。采样过程中,如果不限制标签样本数,强势标签可能因样本数大量减少变为弱势标签,弱势标签也可能因样本数大量增加变为强势标签,这对数据的分布造成破坏,并非数据预处理的初衷。因此,RBS 算法先更新数据集平均样本数,保证弱势标签样本数和强势标签样本数始终分别小于和大于平均样本数,未改变标签对强势和弱势的归属,这在一定程度上维持了原有的数据分布。此外,RBS 过采样产生的新样本,其所属标签集仅含目标标签,RBS 的欠采样仅将目标标签从目标样本的标签集中删除,避免了传统重采样策略样本对其他标签的归属关系和标签原有的依赖关系造成的影响。

c)考虑到算法的欠采样策略客观上并不会减少数据集样本数,从而数据集的样本数越来越多,导致待采样样本数越来越多,这加剧了数据集样本数的增加,进而导致分类算法所需资源越多,耗时越长,算法通过为 meanIR 设定阈值来限制样本规模。实验表明[11],对 $meanIR \leqslant 1.5$ 的数据集进行重采样,对分类算

法性能的提升有限。因此,ML-RBS 设定当数据集 meanIR≤1.5 或最后五次 meanIR 的方差不高于 0.1 后循环终止。

算法 2　随机均衡采样算法 ML-RBS

输入:数据集 mlData,重采样率 P,弱势标签集 minBag,强势标签集 majBag。

输出:重采样后的多标签数据集 returnedData。

```
计算待重采样样本数 samplesToResampling = |mlData| × P;
while samplesToResampling > 0
由式(5)计算 mlData 数据集的平均样本数 meanInstance;
随机选择 L 中的一个标签 label 和该标签的一个样本 y,z 是 y 的副本;
如果 label 是弱势标签,且|label| < meanInstance,执行 6;
设置 z 的 label 的值为 1,其他标签的值为 0,并在 mlData 中添加 z;
如果 label 是强势标签,且|label| > meanInstance,执行 8;
更新 mlData 中 y 的 label 的值为 0;
samplesToResampling - -;
end while
更新 mlData 的平均不平衡度 meanIR;
```

2.3　算法分析

在 ML-RBS 算法中,令多标签数据集样本数为 n,则算法复杂度为 $O(n)$,与算法 1 的 ML-ROS 算法复杂度相同。但对于不平衡度较高的多标签数据集,单独使用过采样方法,样本数会大幅增加,易导致过拟合现象;单独使用欠采样方法,会大量删除强势类样本,造成严重的信息损失。ML-RBS 采用的均衡采样技术,同时对强势类和弱势类数据进行处理,既提高了欠采样对强势类数据的利用率,又减少了非自然增加的弱势类样本数,以一种更均衡的方式减轻了数据集的不平衡度。

综上,ML-RBS 算法提出了平均样本数,一定程度上保持了数据的原始分布;优化了样本复制和删除策略,使不同标签的重采样过程相互独立;结合了过采样和欠采样算法,综合运用强势类和弱势类信息,在不改变算法复杂度的条件下,减少了弱势类信息的冗余和强势类信息的损失。

3　实验分析

3.1　数据集

本文在 Mulan Java 库(http://mulan.sourceforge.net/datasets-mlc.html)所提供的三个多标签数据集上进行实验,各数据集的基本信息如表 1 所示。其中,标签基数(label cardinality,Card)表示平均每个样本有多少个标签,如式(6)所示。

$$\mathrm{Card}=\sum_{i=1}^{m}\frac{|Y_i|}{m} \tag{6}$$

表 1　实验所用数据集及部分信息

数据集	来源领域	样本总数	特征维数	标签基数	标签总数	初始 meanIR
birds	音频	645	260	1.01	19	5.41
CAL500	音乐	502	68	26.04	174	20.58
enron	文本	1 702	1 001	3.38	53	73.95

3.2　实验结果与分析

本文利用 accuracy、macro-FM、micro-FM 以及 AUC 作为算法性能的评价指标[13]。其中,accuracy 为准确度,用于衡量分类正确的比例;macro-FM 和 micro-FM 分别是 F-measure 的宏观平均和微观平均;AUC 是受试者工作特征曲线(ROC)与 X 坐标轴所围成的面积,能有效衡量 MLL 算法性能[14]。

图 2 中,本实验先分别利用 ML-RBS、iML-ROS、ML-ROS 以及传统算法(base),设定重采样率为 0.1、0.2 和 0.25,对数据集进行预处理,其中 iML-ROS 是从 ML-RBS 算法中析出的过采样部分,传统算法是直接对原始数据集进行分类学习;再使用 RAkEL、HOMER、CLR、LP 以及 CC 等 MLL 算法进行分类学习,并进行十折交叉验证,计算评价指标。为客观比较不同方法,每一组数据集都进行五次实验,以其性能指标的平均值来评价系统性能,系统评价指标越高,对应的重采样算法性能越好。

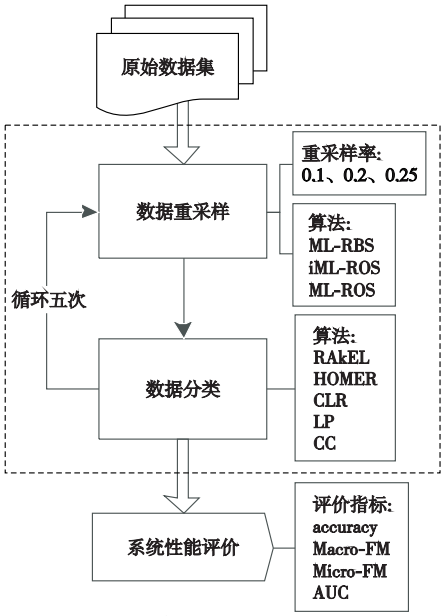
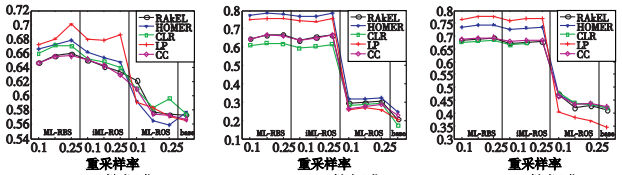


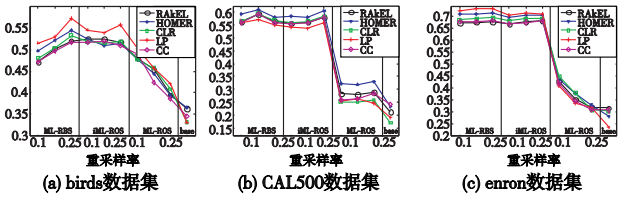
图2　实验流程

实验结果如图 3 ~ 6 所示。图中每一个子图分为四个区域,从左到右依次描述 ML-RBS、iML-ROS、ML-ROS 以及传统算法在三种重采样率下的性能。



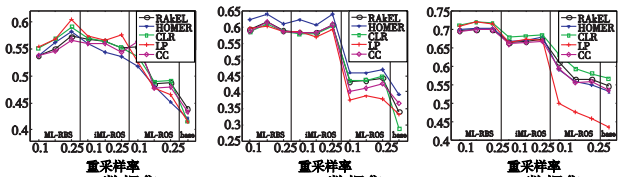
(a) birds数据集　(b) CAL500数据集　(c) enron数据集

图3　不同数据集下重采样算法的accuracy性能



(a) birds数据集　(b) CAL500数据集　(c) enron数据集

图4　不同数据集下重采样算法的macro-FM性能

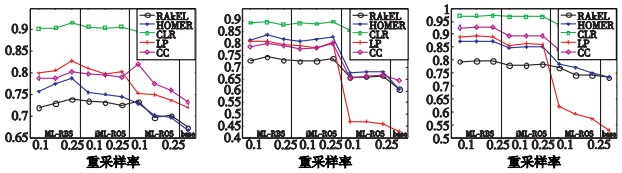


(a) birds数据集　(b) CAL500数据集　(c) enron数据集

图5　不同数据集下重采样算法的micro-FM性能

总的来说,ML-RBS、iML-ROS 和 ML-ROS 都能提升传统算法的分类性能。但是,首先,本文提出的 ML-RBS 及 iML-ROS 的评价指标普遍高于其他算法,特别是对 CAL500 和 enron 这

类不平衡度较高的数据集,算法优势更加明显;然后,ML-RBS 算法性能最好,这不仅体现在该算法能够得到多数性能最优点,而且算法平均性能也普遍优于 iML-ROS;最后,对于 enron 数据集,ML-RBS 和 iML-ROS 算法在不同重采样率下性能相差不大,这说明对于不平衡度较高的数据集,重采样率的变化对算法性能影响较小,本文所提算法性能更加稳定。



(a) birds数据集 (b) CAL500数据集 (c) enron数据集

图6 不同数据集下重采样算法的AUC性能

此外,如图 3 ~5 所示,经 ML-RBS 均衡采样,对于 accuracy、macro-FM 和 micro-FM 评价指标,LP 算法在 birds 和 enron 数据集下性能最好,而 HOMER 算法在 CAL500 数据集下性能最优。图 6 中,在 AUC 评价指标下,CLR 算法在三种数据集下都有最佳性能。

综上所述,本文提出的 ML-RBS 在几种对比算法中具有最好的性能,尤其适用于不平衡度较高的数据集,0.2 和 0.25 是该算法的最佳重采样率。此外,对于重采样后的数据,不同 MLL 算法在不同性能指标下的性能存在差异,实际应用中要根据情况选择合适的分类算法。

4 结束语

数据不平衡问题广泛存在于多标签数据集中。本文提出的 ML-RBS 算法,充分利用了强势类和弱势类的信息,保证了不同标签样本集间的相对独立性,避免样本数的大幅波动,且使得样本数在一定范围内的同时充分降低数据不平衡度。在 birds、CAL500 和 enron 数据集上的实验表明,ML-RBS 在性能上优于其他对比算法,符合预期目标。下一步可从以下方面进行改进:a)提高对不平衡度较低的数据集的处理能力;b)考虑信息熵、样本相似性,进一步优化采样策略;c)算法重采样率的合理选择。

参考文献:

[1] 李凤岐. 基于半监督学习的不平衡数据分类算法与应用[D]. 大连:大连理工大学,2014.

[2] Li Fengqi, Zhang Xuechao, Qiu Tie, *et al*. A pattern query strategy based on semi-supervised machine learning in distributed WSNs[J]. Journal of Information and Computational Science, 2014, 11(18).

[3] 楼晓俊,孙雨轩,刘海涛. 聚类边界过采样不平衡数据分类方法[J]. 浙江大学学报:工学版,2013,47(6):944-950.

[4] Liu Xuying, Li Qianqian, Zhou Zhihua. Learning imbalanced multi-class data with optimal dichotomy weights[C]//Proc of the 13th IEEE International Conference on Data Mining. 2013:478-487.

[5] Fang Ming, Xiao Yuqi, Wang Chongjun, *et al*. Multi-label classification: dealing with imbalance by combining labels[C]//Proc of the 26th International Conference on Tools with Artificial Intelligence. 2014:233-237.

[6] Al-Stouhi S, Reddy C K. Transfer learning for class imbalance problems with inadequate data[J]. Knowledge and Information Systems, 2015, 48(1):1-28.

[7] Tahir M A, Kittler J, Yan Fei. Inverse random under sampling for class imbalance problem and its application to multi-label classification[J]. Pattern Recognition, 2012, 45(10):3738-3750.

[8] Giraldo-Forero A F, Jaramillo-Garzón J A, Ruiz-Munoz J F. Managing imba-lanced data sets in multi-label problems: a case study with the SMOTE algorithm[M]//Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Berlin: Springer, 2013:334-342.

[9] Alejo R, García V, Pacheco-Sánchez J H. An efficient over-sampling approach based on mean square error back-propagation for dealing with the multi-class imbalance problem[J]. Neural Processing Letters, 2015, 42(3):603-617.

[10] Charte F, Rivera A, Del Jesus M J, *et al*. A first approach to deal with imbalance in multi-label datasets[M]//Hybrid Artificial Intelligent Systems. Berlin: Springer, 2013:150-160.

[11] Charte F, Rivera A J, Del Jesus M J, *et al*. Addressing imbalance in multilabel classification: measures and random resampling algorithms[J]. Neurocomputing, 2015, 163:3-16.

[12] Japkowicz N, Stephen S. The class imbalance problem: a systematic study[J]. Intelligent Data Analysis, 2002, 6(5):429-449.

[13] Madjarov G, Kocev D, Gjorgjevikj D, *et al*. An extensive experimental comparison of methods for multi-label learning[J]. Pattern Recognition, 2012, 45(9):3084-3104.

[14] Huang Jin, Ling C X. Using AUC and accuracy in evaluating learning algorithms[J]. IEEE Trans on knowledge and Data Engineering, 2005, 17(3):299-310.

(上接第 2900 页)

[3] Fournier-Viger P, Wu Chengwei, Zida S, *et al*. FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning[M]//Foundations of Intelligent Systems. [S. l.]: Springer International Publishing, 2014:83-92.

[4] Zida S, Fournier-Viger P, Lin Chunwei, *et al*. EFIM: a highly efficient algorithm for high-utility itemset mining[C]//Advances in Artificial Intelligence and Soft Computing. [S. l.]: Springer International Publishing, 2015:530-546.

[5] Wu Chengwei, Shie B E, Tseng V S, *et al*. Mining top-*k* high utility itemsets[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012:78-86.

[6] Ryang H, Yun U. Top-*k* high utility pattern mining with effective threshold raising strategies[J]. Knowledge-Based Systems, 2015, 76(1):109-126.

[7] Tseng V S, Wu Chengwei, Fournier-Viger P, *et al*. Efficient algorithms for mining top-*k* high utility itemsets[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(1):54-67.

[8] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.

[9] Li Haoyuan, Wang Yi, Zhang Dong, *et al*. PFP: parallel FP-growth for query recommendation[C]//Proc of ACM Conference on Recommender Systems. New York: ACM Press, 2008:107-114.

[10] 陈光鹏,杨育彬,高阳,等. 一种基于 MapReduce 的频繁闭项集挖掘算法[J]. 模式识别与人工智能,2012,25(2):220-224.

[11] 杨勇,高松松. 基于 MapReduce 的关联规则并行增量更新算法[J]. 重庆邮电大学学报:自然科学版,2014,26(5):670-678.

[12] 唐颖峰,陈世平. 一种基于后缀项表的并行闭频繁项集挖掘算法[J]. 计算机应用研究,2014,31(2):373-377.