文章编号：2096-1472(2019)-01-08-04
DOI:10.19644/j.cnki.issn2096-1472.2019.01.002

基于RAKEL算法的商品评论多标签分类研究与实现

梁睿博，王思远，李 壮，刘亚松
(东北大学计算机科学与工程学院，辽宁 沈阳 110819)

摘 要：商品通常包含多个属性维度，准确找到商品评论中涉及的属性维度是文本挖掘工作的基础。RAKEL算法是多标签分类中问题转换思路的一种实现。在以往的工作中，由于子标签集合的随机性，没有充分发现和考虑标签之间的相关性，导致分类精度不高。为此，提出了改进的FI-RAKEL算法。首先通过FP-Growth算法得到标签的频繁项集，再从频繁项集和原始标签集合中选择标签构成新的标签子集，以此充分利用标签相关性训练基分类器。实验证明，改进的FI-RAKEL算法具有更好的评论文本多标签分类性能。

关键词：多标签分类；RAKEL；频繁项集；标签相关性
中图分类号：TP391 **文献标识码**：A

Research and Implementation of RAKEL Algorithm Based Multi-Label Classification for Online Commodity Reviews

LIANG Ruibo,WANG Siyuan,LI Zhuang,LIU Yasong
(*School of Computer Science and Engineering*,*Northeastern University*,*Shenyang* 110819,*China*)

Abstract: Generally,there are multiple attribute-dimensions to describe a commodity.It is the foundation of text mining to accurately find the attribute-dimensions involved in commodity reviews.The Random K-Labelsets (RAKEL) is an accomplishment of problem transformation in multi-label classification.However,due to the randomness of sub-labelset and the lack of investigating into the relationship among labels,the classification accuracy of RAKEL is not high.Hence,an improved RAKEL algorithm (FI-RAKEL) is proposed.Firstly,the item-frequency sets of labels are obtained through the FP-Growth algorithm.Then,labels are selected from the item-frequency sets and the original label set respectively to generate a new k-labelset and it is used to train the corresponding classifier based on correlation among labels.The experiment result shows that the proposed FI-RAKEL algorithm brings higher classification accuracy for multiple-labeled reviews.

Keywords:multi-label classification;RAKEL;item-frequency set;label correlation

1 引言(Introduction)

近些年，网购成为了人们日常消费的主要方式。由此，各大电商平台上积累了海量的用户购物评论数据，其中蕴藏着巨大的商业价值。一方面，用户评论是企业和商家了解市场反馈的重要渠道；同时，对于消费者而言，参考其他人发表的评论也有助于快速地选择理想的商品。通常，一种商品会包含多个属性维度，用户针对某个商品发表的评论也会涉及商品的多个方面。因此，对商品评论进行文本挖掘时，准确找到评论中涉及的属性维度是整个文本挖掘工作的基础。针对商品评论数据集，多标签分类算法是首要考虑的问题。

多标签分类算法主要研究当样本同时具有多个类别标记时，如何构建分类器，准确预测未知样本的标签集合[1]。本文首先从京东商城等电商平台按品类获取了商品评论，并对这些评论进行人工标注。按照标签对商品评论文本进行统计后发现，一些标签之间具有较高的相关性，例如，表1列举的洗发水商品的评论R1-R6。从表1中可以看出，“快递”和“购物渠道”这两个标签在同一条用户发表的评论文本中共现(被同时提及)的比例较高，我们可以认为这两个标签具有一定的相关性。导致这一现象的原因是，当“购物渠道”为电商平台时，用户必然会接受快递服务，因此两者的共现概率较高。而在实际应用中，标签之间是存在一定联系的。本文以标签相关性为基础，参考近年来基于标签相关性的多标签分类算法，提出了基于频繁项集的改进RAKEL算法FI-RAKEL。首先，通过频繁项集挖掘标签之间的关联关系，选取频繁项集的元素作为RAKEL算法的标签子集，从而利用标签间的相关性提高预测分类的精确度和整体性能。

基金项目：本文受the National Key R&D Program of China under grant(2018YFB1004700)资助.

2 相关工作(Related work)

多标签学习的研究，起源于2000年的Schapire等提出的基于boost方法的文本多分类，著名的学者Tsoumakas、Jesse Read等从事过相关研究。解决多标签分类问题主要有两种思路：算法适应和问题转换[2-4]。问题转换法通过对样本集合进行分解，达到把多标签学习问题转换为多个单标签学习问题的目的。该方法具有简化性，并且在大多数据集上应用良好[5]，也是本文主要采用的方法。

问题转换算法中经典的方法有BR(Binary Relevance)、CC(Classifier Chains)和LP(Label Powset)等。其中BR算法是一阶方法，将多标签学习问题分解为多个独立的二元分类问题，该方法完全忽略了标签之间的潜在相关性。在BR的基础上，Jesse Read等[6]人提出了Classifier Chains算法，将多标签学习问题转化为二元分类问题链，链中的后续二元分类器基于前面的分类器进行预测[7]。分类器链具有开发标签相关性的优点，但由于其链接属性而无法实现并行。研究者还根据BR、CC等思想提出了Ensemble的框架[8]，提出了EBR、ECC等算法，这些算法也表现出了很好的性能。

表1 京东商城洗发水产品的评论

Tab.1 Examples of product's reviews

标签	评论
快递、购物渠道	R1：非常喜欢，发货速度快，一直在京东买东西，相信京东
	R2：我是京东铁粉，信赖，保真，服务好，送货快
产品质量、品牌	R3：一直在用，认定了这品牌，好用好用
	R4：洗发露用着不错，绝对正品，老牌子一直都在使用
购物渠道、价格	R5：很便宜，京东东西信得过
	R6：便宜实惠，价格爆炸，服务一流，真的是大爱京东

另一种常见的问题转换思路是创建新标记，其中LP算法是将每个多标签实例的所属标签联合起来创建新的标签，但是这样做会大大增加标签数量，增加计算开销。后来的研究者们在LP思想的基础上提出了Pruned Problem Transformation(PPT)算法[9]和Random k-labelsets(RAKEL)算法，以及一些RAKEL改进算法[10-13]。RAKEL的基本思想是将多标签学习问题转化为多类分类问题的集合，从标签集合中随机选出小部分标签子集，在这个子集的多分类分类器上引入Label Powerset(LP)技术。RAKEL是一个高阶方法，其中标签相关度由k-labelsets的大小来控制，避免了LP的缺陷。但是正因为标签子集的随机选择，对标签之间的相互联系考虑不足，从而导致分类的精确度不高。对此，研究者分别做出了不同的改进。文献[10]提出的RAKEL改进算法LC-RAKEL，核心思想是通过聚类来选取标签子集。对随机选择的k个标签进行聚类，从每个聚类的标签簇中选取一个标签形成标签子集，通过训练可以得到分类精度较高的子分类器。但当标签数目较少并且相关度较高时，聚类效果不理想。文献[13]提出一种基于成对标签的RAKEL改进算法PwRAKEL。该方法考察任两个标签的共现性，利用生成的共现矩阵选择共现度高的成对标签加入标签子集，提高标签之间的相关关系来提升子分类器的模型预测精度。这种方法只考虑了每两个标签间的相关关系，没有将更多标签相互关联的情形充分利用。

还有一些学者进行了基于频繁项集的多标签分类算法改进[14,15]。文献[15]提出了一种基于频繁项集的多标签文本分类算法MLFI，利用FP-growth算法挖掘类别之间的频繁项集，同时为每个类计算类标准向量和相似度阈值，如果文本与类标准向量的相似度大于相应阈值则归到相应的类别，在分类结束后利用挖掘到的类别之间的关联规则对分类结果进行校验。该方法主要针对文档进行类别划分，不适用于商品评论的短文本多标签分类问题。

基于以上的相关工作，本文针对商品评论提出一种改进的多标签分类方法FI-RAKEL。首先对标签之间相关性进行频繁模式挖掘，选取频繁项集作为RAKEL算法的标签子集，充分利用标签相关性来训练子分类器，提升分类器的预测分类精度，实现RAKEL算法的改进。

3 基于RAKEL算法的商品评论多标签分类算法(Algorithm of RAKEL algorithm based multi-label classification for online commodity reviews)

Tsoumakas等提出的Random k-Labelsets将集成学习与LP结合，将原始大标签集分成小标签集，使用LP技术训练相应的分类器。对于新实例的多标签分类，每个分类器为每个标签提供二元决策，随后计算每个标签的平均决策，如果平均值大于用户指定的阈值则输出最终的肯定决策。RAKEL算法对于标签子集的择是随机的，没有充分利用标签相关性，对最终的分类结果产生巨大影响。本文从这一角度展开，利用标签间的相关性来确定标签间的关系，提出了FI-RAKEL算法。

表2 不同MinSup下的频繁项集

Tab.2 Frequent itemsets with different MinSup

最小支持度MinSup	频繁项集
0.15	{<香味>,<品牌>,<产品功效>,<质量>,<质量,购物渠道>,<性价比>,<性价比,购物渠道>,<快递,性价比>,<购物渠道>,<购物渠道，快递>,<快递>}
0.2	{<香味>,<品牌>,<产品功效>,<质量>,<质量,购物渠道>,<性价比>,<购物渠道>,<购物渠道，快递>,<快递>}
0.25	{<品牌>,<产品功效>,<质量>,<质量,购物渠道>,<性价比>,<购物渠道>,<购物渠道，快递>,<快递>}
0.3	{<产品功效>,<质量>,<质量,购物渠道>,<性价比>,<购物渠道>,<购物渠道，快递>，<快递>}
0.32	{<质量>,<质量,购物渠道>,<性价比>,<购物渠道>,<购物渠道，快递>,<快递>}}
0.35	{<质量>,<质量,购物渠道>,<性价比>,<购物渠道>,<快递>}

首先采用FP-Growth算法实现标签间的关联关系的挖掘，得到标签的频繁项集。FP-Growth算法对数据集进行扫描，得到标签的频繁项列表并对频繁项进行支持度递减的排序，我们设定最小支持度*MinSup*，认为小于*MinSup*值的项为非频繁项，从而去除其中的非频繁项；第二次扫描构造一棵FP-Tree[15]，FP-Tree是从上至下发散的层次树，越是上层的点表示出现越频繁。从FP-Tree中提取频繁项，得到标签的频繁项集。不同*MinSup*下频繁项集输出结果如表2所示，在FI-RAKEL算法中频繁项集的项数n的取值范围是$1 \le n \le k$。基于得到的频繁项集构造RAKEL算法的标签子集，从频繁项集中选取一个频繁项和原始标签集合中的部分标签作为标签

子集来训练分类器，最后实现预测分类。

为了更好地描述算法，首先引入一些相关定义：令$L=\{\lambda_1,\lambda_2,\lambda_3,\cdots\}$代表多标签分类域中的标签集合，大小为$|L|$。集合$Y\subseteq L$并且$k=|Y|$，则称$Y$为$k-labelsets$。在此，使用$L^k$表示$L$上所有不同$k-labelsets$的集合，$D$为训练数据集。

首先在数据集上通过FP-Growth算法生成标签的频繁项集$F=\{f_1,f_2,f_3,\cdots\}$，然后迭代构造m个LP分类器的集合。在每次迭代中即$i=1,\cdots,m$，分别从频繁项集F中拿出第i个频繁项f_i和L中随机选择$(k-|f_i|)$个不重复的标签构成标签子集，训练子分类器h_i。对于新实例x，每个模型h_i为相应k标签集Y_i中的每个标签λ_j提供二元决策$h_i(x,\lambda_j)$，每个标签λ_j的平均决策，如果平均值大于用户指定的阈值t，则输出最终的肯定决策。FI-RAKEL算法模型训练过程如算法1。

算法1 FI-RAKEL模型训练算法

输入：模型数目m，频繁项集最小支持度$MinSup$，子集大小k，标签集合L，训练集D，空集R

输出：LP分类器h_i，相应的$k-labelset\ \ Y_i$

$F=\{f_1,f_2,f_3,\cdots\}\leftarrow$FP-Growth($MinSup$，$D$)

for $i\leftarrow 1$ to $\min(m,L^k)$

if $(i\le|F_n|)$

$Y_i\leftarrow$i-th frequent item f_i of labels selected from $F+(k-|f_i|)$ labels randomly selected from L not in f_i

else

$Y_i\leftarrow$a k-labelset randomly selected from L

endif

if (Y_i not in R)

put Y_i in R

else

repeat 3 to 12

endif

train an LP classifier $h_i:X\rightarrow P(Y_i)$ on D

endfor

4 实验与结果分析(Experiment and analysis)

4.1 实验方法

4.1.1 实验数据集

本文以京东商城和天猫商城作为数据源，获取其中洗发水商品评论。选取50万条洗发水商品评论，进行人工属性维度的标注，以领域专家指定的10个商品维度作为标签，包括“质量”“产品功效”“香味”“质地”“包装”“性价比”“品牌”“购物渠道”“快递”和“方便性”，以此作为实验的训练集和测试集。在计算测试集相对训练集占比时，从数据集中随机抽取数据作为测试集，其余作为训练集，得到测试集大小与训练集大小比例。

4.1.2 评价指标

根据文献[2]，本文从几个方面对多标签分类算法进行评估分析：Subset-Accuracy、Recall和F-Measure，定义如式(1)—式(3)。

①Subset-Accuracy

$$SubsetAccuracy=\frac{1}{p}\sum_{i=1}^{p}I(Z_i=Y_i) \tag{1}$$

②Recall

$$Racall=\frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i\bigcap Z_i|}{|Y_i|} \tag{2}$$

③F-Measure

$$F-Measure=\frac{1}{p}\sum_{i=1}^{p}\frac{2|Y_i\bigcap Z_i|}{|Z_i|+|Y_i|} \tag{3}$$

其中，p表示实验的测试集大小，式(1)中I的取值分别为$I(true)=1$和$I(false)=0$，Subset-Accuracy评估正确分类例子的分数，即预测标签集合与实况标签集合的相同程度。式(3)中F-Measure是衡量算法整体性能的基本指标。

4.2 实验结果和分析

基于4.1.1中的商品评论数据集，本文对FI-RAKEL算法进行验证并与其他RAKEL改进算法进行对比实验，参考4.1.2中的评价指标。

首先针对本文提出的FI-RAKEL算法进行实验，FI-RAKEL算法采用的基分类器为LP分类器，LP的基分类器选择SVM分类算法，标签子集大小为$k=7$，模型个数$m=2|L|$，预测分类中阈值$t=0.5$。我们选取不同的频繁项集最小支持度$MinSup$进行对比实验，其中训练集相对数据集的占比为0.8，使用Subset-Accuracy作为评价指标。同时，我们在所有实验过程中选择三次交叉验证，即测试集相对训练集占比一定时，进行三次测试集的随机抽取，得到多个结果的平均值。实验结果如图1所示。

从图1我们可以看出当$MinSup$取值范围在$[0.30,0.35]$时，FI-RAKEL算法分类精度最高。我们又以0.01为间距进行实验，发现$MinSup=0.32$时算法取得最好的分类精确度。随后我们选择REKAL算法和文献[13]提出的PwRAKEL与本文算法进行对比实验，基分类器为LP分类器，LP的基分类器同样使用SVM分类算法，算法取标签子集大小为$k=7$，模型个数$m=2|L|$，预测分类中阈值$t=0.5$。对FI-RAKEL算法最小支持度取值为$MinSup=0.32$。实验结果如图2所示。

通过分析实验结果我们发现，当测试集相对训练集占比增加时算法性能趋优。如图2所示，当测试集与训练集比例为0.9时，FI-RAKEL算法在分类准确率、召回率和F值等评价指标上有非常明显的优势。相较于RAKEL算法，FI-RAKEL有更高的分类准确率召回率和F1值，这说明本文提出的基于频繁项集构建标签子集的思想能够有效利用标签相关性提升子分类器的训练和预测分类精度。

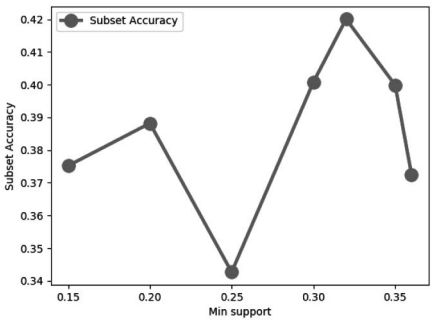
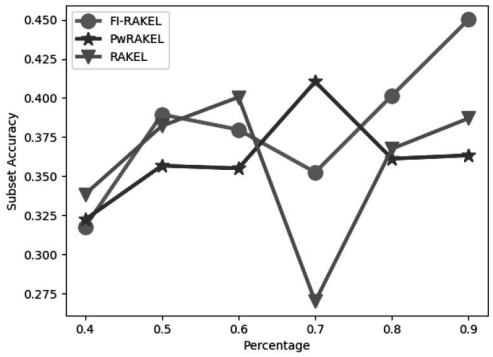
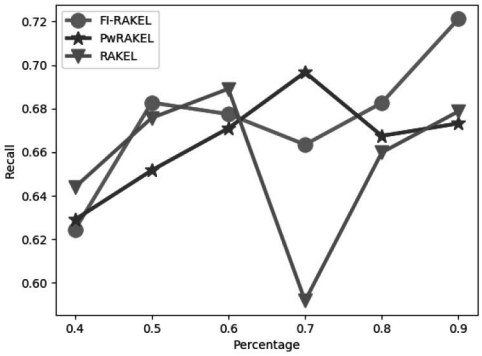


图1 FI-RAKEL在不同MinSup下评价

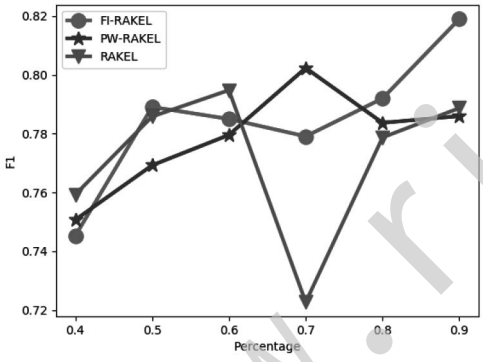
Fig.1 Evaluation of FI-RAKEL with different MinSup



(a) Sub-Accuracy



(b) Recall



(c) F-Measure

图2 各算法的性能对比

Fig.2 Performance comparison of algorithms

5 结论(Conclusion)

在多标签分类问题上，RAKEL算法由于标签子集选择的随机性，没有充分利用标签间的相关性来改善算法的性能。针对商品评论文本挖掘这一领域，本文提出了一种基于频繁项集的RAKEL改进算法FI-RAKEL，并与RAKEL算法和PwRAKEL算法进行对比，实验结果表明FI-RAKEL算法性能更好，具有更高的分类准确性、召回率和F值。虽然本文算法在性能上有些优势，但在实现本文过程中发现生成了大量新标签组合，其中有些标签组合在结果中出现频率较低造成了资源浪费，因此如何更合理利用标签相关性，过滤低频率标签组合并找到隐藏相关标签，减少资源浪费和运行时间是下一步值得思考的。

参考文献(References)

[1] G.Tsoumakas,I.Katakis,and I.Vlahavas,Random k-labelsets for multi-label classification,IEEE Trans.Knowl.Data Eng.,2011,23(7):1079-1089.

[2] Padmanabhan Divya,Bhat Satyanath,Shevade Shirish,Narahari Y.Topic Model Based Multi-Label Classification.2016 IEEE 28th International Conference on Tools with Artificial Intelligence,Nov 2016:996-1003.

[3] Huang Jun,Li Guorong,Huang Qingming,et al.Learning Label Specific Features for Multi-label Classification.2015 IEEE International Conference on Data Mining,2015(11):181-190.

[4] 张洛阳,毛嘉莉,刘斌,等.基于贝叶斯模型的多标签分类算法[J].计算机应用,2016(1):52-56;71.

[5] 徐婧扬.多标签分类算法研究及其应用[D].山东大学,2017.

[6] Read Jesse,Martino Luca,Luengo.Efficient monte carlo methods for multi-dimensional learning with classifier chains.Pattern Recognition,March 2014,47(3):1535-1546.

[7] Yu Zhilou,Hao Hong,Zhang Weipin.A Classifier Chain Algorithm with K-means for Multi-label Classification on Clouds.Journal of Signal Processing Systems,2017,86(2):337-346.

[8] Rokach Lior,Schclar Alon,Itach Ehud.Ensemble methods for multi-label classification.Expert Systems With Applications,November 2014(41):7507-7523.

[9] Read J.A pruned problem transformation method for multi-label classification[C].Proceeding of New Zealand Computer Science Research Student Conference.Christchurch:Canterbury University,2008:143-150.

[10] 金永贤,张微微,周恩波.一种改进的RAKEL多标签分类算法[J].浙江师范大学学报(自然科学版),2016,39(4):386-391.

[11] Wu Yu-Ping,Lin Hsuan-Tien.Progressive random k-labelsets for cost-sensitive multi-label classification.Machine Learning,2017,106(5):671-694.

[12] Osojnik,Aljaž,Panov.Multi-label classification via multi-target regression on data streams.Machine Learning,2017,106(6):745-770.

[13] 周恩波,叶荣华,张微微,等.一种基于成对标签的Rakel算法改进[J].计算机与现代化,2016(3):16-18;23.

[14] 吕小勇,石洪波.基于频繁项集的多标签文本分类算法[J].计算机工程,2010(15):83-85.

[15] Jiawei Han,Jian Pei,Yiwen Yin.Mining Frequent Patterns without Candidate Generation:A Frequent-Pattern Tree Approach[J].Data Mining and Knowledge Discovery,2004(8):53-87.

作者简介：

梁睿博(1994-)，女，硕士生.研究领域：文本挖掘，机器学习.
王思远(1995-)，男，硕士生.研究领域：深度学习，文本挖掘.
李　壮(1993-)，女，硕士生.研究领域：文本挖掘，机器学习.
刘亚松(1992-)，男，硕士生.研究领域：机器学习，知识图谱.