一种基于标签相关性的多标签分类算法*

王 霄[1]，周李威[1]，陈 耿[2]，朱玉全[1]

(1. 江苏大学 计算机科学与通信工程学院，江苏 镇江 212013；2. 南京审计学院 信息科学学院，南京 211815)

摘 要： 针对基于概率统计的 ML-kNN 算法只能对每个独立的标签进行分析，忽略了真实世界中标签间的相关性，提出了一种联系标签相关性的 ML-kNN 算法(S-ML-kNN)。该方法对训练集进行扩展，并按照标签间的二阶组合来构造新的标签，融合了标签之间的相关性。实验结果表明，S-ML-kNN 算法优于 ML-kNN 算法。

关键词： 多标签；标签相关性；kNN；二阶

中图分类号： TP391；TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2014)09-2609-04

doi：10.3969/j.issn.1001-3695.2014.09.011

Correlation label-based multi-label classification algorithm

WANG Xiao[1], ZHOU Li-wei[1], CHEN Geng[2], ZHU Yu-quan[1]

(1. *School of Computer Science & Telecommunications Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*; 2. *School of Information Science, Nanjing Audit University, Nanjing* 211815, *China*)

Abstract: The only ML-kNN algorithm based on probability and statistics for each individual tag analysis, ignoring the correlation between the tags in the real world, this paper proposed a ML-kNN algorithm with the label correlation(S-ML-kNN), the method extended the training set and follow the label between the second combination to construct a new label, the integration of the correlation between labels. Experimental results show that, S-ML-kNN algorithm outperforms ML-kNN algorithm.

Key words: multi-label; label correlation; kNN; second order

与单标签学习相比，多标签学习是一种更符合真实世界客观规律的方法，尤其在文本分类[1~3]、图像分类[4,5]、生物基因功能分类[6]等领域有着广泛的应用。

对于多标签分类问题，目前主要的解决途径[7]有问题转换和算法适应。问题转换法的主要思想是通过对已知的训练集进行处理，将多标签学习问题转换为其他已知的学习问题进行求解。BR(binary relevance)[8]方法是一种典型的基于数据分解的方法，它把每一个标签的预测视为一个独立的单分类问题，并为每一个标签训练一个独立的分类器，用全部的训练数据对每个分类器进行训练。这种方法简便易行，但忽略了标签之间的相互关系，预测结果也往往难以令人满意。CLR[9]方法加入了人工校准标签来区分相关标签和不相关标签，但是当数据集中的类别标签很多的时候，这种方法构造出的子分类器过多，从而增加了算法的复杂度，也会对预测结果产生很大影响。与之前的方法相比较，RAkEL[10]方法考虑到了标签之间的依赖关系，且弥补了 LP 方法可能产生偏斜数据的不足。然而这种方法想要达到最佳效果需要大量数据集，必须要对输入参数如子集大小、阈值等进行内部交叉检验(internal cross validation)，而在训练样本不足的情况下很难找到最优化的参数。算法适应法的主要思想是通过对常用监督学习算法进行改进，将其直接运用到多标签学习的问题上来。C4.5 算法可以从数据集中学习到一些精确而有意义的多标签分类规则，但却不能解决完全的分类问题。BP-MLL(back-propagation for multi-label learning)[11]反向传播多标签学习法引入一个新的误差函数，引入 ranking loss 因素，减少运算时间，但是却大大增加了计算复杂度。ML-kNN[12]算法是一种简单且非常有效的解决多标签问题的方法，它利用最大化后验原则来确定待预测样本的标签集。然而，由于它仅针对每一个独立标签来统计其在近邻中被包含的数量，却忽略了各个标签之间可能存在的相关性。

真实世界中，标签与标签之间往往不是相互独立，而是有一定联系的。因此，在多标签学习中，可以利用标签之间的相关性来辅助解决问题。例如，如果一幅图像包含标签“武器”和“军人”，那么该图片包含标签“军队”的可能性就会比较大。因此，如何充分利用标签间的相关性是构造具有强泛化能力多标签学习系统的关键。然而上述方法均未能很好地利用标签之间潜在的语义相关性和共现性知识。为了解决存在的这一问题，本文利用从文本检索中受启发而得到的词与词的共现概率来对 ML-kNN 算法进行改进。S-ML-kNN 算法把标签相关性强弱融合到原始的 ML-kNN 中，由比较最终的后验概率来判断标签的包含情况。

1 相关工作

1.1 多标签问题的定义

设 $X=\{x_1,x_2,\cdots,x_m\}$ 代表示例空间，$L=\{L_1,L_2,\cdots,L_q\}$ 代表所有的标签集合，$Y=\{y_1,y_2,\cdots,y_m\}$ 代表标签空间，则学习系统的任务是从训练集 $\{(x_i,y_i)\mid 1\leqslant i\leqslant m\}$ 中学得函数 $f:X\rightarrow Y$，其中 $x_i\in X$ 为一个示例，$y_i\in Y$ 为示例 x_i 所属的类别标签，且 y_i 为标签集合 L 的一个子集。当学习对象的类别标签唯一，即 x_i 与 y_i 可以一一对应起来时，这种传统的监督学习框

收稿日期： 2013-10-09；**修回日期：** 2013-11-11 **基金项目：** 国家自然科学基金资助项目(71271117)；江苏省科技型企业技术创新资金项目(BC2012331)

作者简介： 王霄(1989-)，男，江苏徐州人，硕士研究生，主要研究方向为模式识别、人工智能、数据挖掘等(821074860@qq.com)；周李威(1989-)，男，江苏盐城人，硕士，主要研究方向为模式识别、数据挖掘等；陈耿(1965-)，男，教授，博士，主要研究方向为数据挖掘、审计风险管理等.

架已经取得了很大的成功。然而事实上,真实世界中的对象往往并不是只具有唯一的语义,很有可能具有多义性,因此,便引出多标签学习框架。在该框架下,每个对象由一个示例描述,此时该示例具有多个而不是唯一的类别标签,学习的目标则是将所有合适的类别标签赋予未知示例。

1.2 求解策略

基于考察标签相关性的不同方式,已有的多标签学习问题求解策略大致可以分为如下三类[13]:

a) 一阶策略。依次考察每个标签,将多标签学习问题分解为 q 个独立的二类分类问题。该方法实现简单但泛化性能较低。

b) 二阶策略。考察标签两两之间的相关性。该方法在一定程度上考虑到了标签间的相关性,但是不能包含所有的标签相关情况。

c) 高阶策略。考察高阶的标签相关性。考虑较为全面,但是计算复杂度很高,难以处理大规模学习问题。

1.3 ML-kNN 算法简述

张敏灵等人提出的 ML-kNN 算法的基本思想是采用 k 近邻(k-nearest neighbors)分类准则,统计这 k 个近邻示例包含标签的信息,通过最大化后验概率的方式推理未见示例的标签集合。参照 1.1 节介绍,X 是特征空间,Y 表示标签空间。

把 Y_i 表示成向量的形式 $Y_i(l)$,$l=1,2,\cdots,q$ 表示的是 x_i 的第 l 个类标签分量,当 x_i 包含第 l 个标签时,$Y_i(l)$ 的值就是 1,否则为 0。$N(x_i)$ 表示训练数据集中 x_i 的 k 个邻居的标签。$C_t(i)$ 为示例 t 的 K 个近邻示例中恰好有 $C_t(i)$ 个包含标签 L_i 的近邻数(满足 $0\leqslant C_t(i)\leqslant K$)。$H_b^l$ 为示例 t 包含($b=1$)和不包含($b=0$)标签 L_l 的假设。E_j^l 为示例 t 的 K 个近邻示例中恰好有 j 个包含标签 L_l 的事件。

该算法引入贝叶斯概率,将 kNN 的分类函数修改为

$$Y_t(l)=\underset{b\in\{0,1\}}{\operatorname{argmax}} P(H_b^l)P(E_j^l|H_b^l)\quad 1\leqslant l\leqslant q \tag{1}$$

即通过比较包含和不包含标签 L_l 的大小来最终确认示例 t 能否包含标签 L_l。

对于数据集中的每一个独立标签 L_l,其对应的先验概率 $P(H_b^l)$ 可以由式(2)得到。

$$P(H_1^l)=(s+\sum_{i=1}^{m}Y_i(l))/(s\times 2+m);P(H_0^l)=1-P(H_1^l) \tag{2}$$

条件概率 $P(E_j^l|H_b^l)$ 则可以通过式(3)和(4)得到。

$$P(E_j^l|H_1^l)=(s+c[j])/(s\times(K+1)+\sum_{p=0}^{k}c[p]) \tag{3}$$

$$P(E_j^l|H_0^l)=(s+c'[j])/(s\times(K+1)+\sum_{p=0}^{k}c'[p]) \tag{4}$$

由于 ML-kNN 算法采用的是一阶策略来求解多标签学习问题,即在模型构建过程中忽略标记之间的相互影响,因此其计算结果也是基于各个标签均不相关的情况。然而在实际情况下,标签与标签之间不可避免地会产生相关联性,这是个不容忽视的问题,所以本文据此作出改进。

2 一种基于标签相关性的 ML-kNN 多标签分类算法

2.1 算法的基本思路

本文采用二阶策略,对于所有的标签集合 $L=\{L_1,L_2,\cdots,L_q\}$,任取两个标签考察其相关性,共有 $q(q-1)/2$ 种组合方式,再加上原来的 q 个标签,共计有 $q+q(q-1)/2$ 个不同的标签,即新的标签集合 L_{new} 的势 $|L_{new}|=q+q(q-1)/2$。对于训练集中任一示例 x_i,其对应的 0/1 标注 L_1-L_q 与扩展前一致,$L_{q+1}-L_{q+q(q-1)/2}$ 则要依照原先的标注进行修正,同时包含标记 L_i 和 L_j($1\leqslant i\leqslant q$,$1\leqslant j\leqslant q$ 且 $i\neq j$)的标上 1,否则标上 0。

在文本检索中,文献[14]提出的自动局部分析利用词与词的共现模式,能够有效地找到与查询项语义相关的词进行查询扩展。受此启发,若将训练集中每幅图像看成包含标注词的文档,则可用类似的方法进行标注词之间语义相关性的度量。对于训练集中的已知分类情况的示例,可以获得标注词与示例构成的共现矩阵 M,如表 1 所示。

表 1 共现矩阵 M

		x_1	x_2	x_3	x_4	…	x_n
ω_1	L_1	0	1	0	1	…	1
ω_2	L_2	1	0	1	0	…	1
ω_3	L_3	1	0	1	1		0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
ω_q	L_q	1	1	0	0	…	1

在表 1 中,x_i($1\leqslant i\leqslant m$)表示训练集中的第 i 个示例,ω_j($1\leqslant j\leqslant q$)表示矩阵 M 第 j 行的所有数据,通过公式

$$C_{\omega u,\omega v}=\sum M_{ui}\times M_{vi} \tag{5}$$

$$S_{\omega u,\omega v}=\frac{C_{\omega u,\omega v}}{C_{\omega u,\omega u}+C_{\omega v,\omega v}-C_{\omega u,\omega v}} \tag{6}$$

进行归一化处理,可得到 ω_u 和 ω_v 共同出现的频率 S_{ω_u,ω_v},即标签 L_u 和 L_v 的共现频率。这样处理之后得到的结果 S 构成了一个对称矩阵,且对角元素均为 1(表示的是标签 L_i 与其本身的共现概率为 1,这是必然事件)。在改进的 ML-kNN 算法中,对于每一个标签,都要考虑它与其他标签的共现概率(该标签本身除外)。通过上述方法,不仅可以从训练数据集中获取更多信息,还能衡量出任意两个标签间的相关性大小。

对于一个测试示例 t_i,将得到的共现概率 $Q1_{ij}$ 应用到式(7)中以计算 $YY1(i)$。利用 $Q0_{ij}=1-Q1_{ij}$。将结果代入式(8)计算 $YY0(i)$。

$$YY1(i)=\sum Q1_{ij}\times PEH1(ij,d) \tag{7}$$

$$YY0(i)=\sum Q0_{ij}\times PEH0(ij,d) \tag{8}$$

设定参数 α,满足 $0\leqslant\alpha\leqslant 1$,将上述计算结果与 ML-kNN 中计算得到的 $Y1(i)$ 和 $Y0(i)$ 结合起来,经过式(9)(10)计算得到 $y1(i)$ 和 $y0(i)$,通过比较两者大小来确定测试实例是否包含标签 Li。

$$y1(i)=\alpha\times Y1(i)+(1-\alpha)\times YY1(i) \tag{9}$$

$$y0(i)=\alpha\times Y0(i)+(1-\alpha)\times YY0(i) \tag{10}$$

2.2 算法描述

根据 2.1 节对标签之间相关系数计算的阐述和上述有关标签组合的定义,本文提出了一种联系标签相关性的 ML-kNN 多标签分类算法,具体描述如算法 1 所示。

算法 1 基于标签相关性的 ML-kNN 多标签分类算法 S-ML-kNN。

输入: 训练数据集 X 和测试数据集 T,所有的标签集合 L,标签空间 Y,设置近邻参数 K,设置平滑参数 s,设置系数 α。

输出: 测试数据集对应的标签集。

```
1 for i∈{1,2,…,q} do
2   for j∈{1,2,…,m} do //m 为训练集中示例个数
3     A(i)=0//A(i) 用来统计训练集中包含标签 Li 的示例个数
4     if (Train_ij)==1) then A(i)=A(i)+1 //Train 为训练数据集
5   P0(i)=1-P1(i)
6   end
7 end
8 for i∈{q+1,q+2,…,q(q-1)/2} do
```

9　分别找到与标签 L_i 配对的标签,在 S_n 中找到 $Sn_{ij}(i\neq j)$,记为 $Q1_{ij}$

10　$Q0_{ij}=1-Q1_{ij}$

11 end

12 对训练集中的每一个实例 x_i,先找到它的 K 个近邻,记为 $N(x_i)$,$i\in\{1,2,\cdots,m\}$

13 for　$l\in Y$ do

14　for　$j\in\{1,2,\cdots,K+1\}$ do

15　$C1[j]=0$; $C2[j]=0$; //$C1$、$C2$ 均初始化为 0,用来储存 K 个近邻中分别包含和不包含标签 Li 的个数

16　end

17　for　$i\in\{1,2,\cdots,m\}$ do

18　　令 $\sigma=Xj$ 近邻中恰好包含标签 Li 的个数

19　　if($Train_{ij}==1$) then　$C1[\sigma]=C1[\sigma]+1$;

20　　else　$C2[\sigma]=C2[\sigma]+1$;

21　end

22　for　$j\in\{1,2,\cdots,K+1\}$ do

23　　$PEH1=(s+C1[j])/(s*(K+1)+\sum_{p=1}^{k+1}C1[p])$

24　　$PEH0=(s+C2[j])/(s*(K+1)+\sum_{p=0}^{k+1}C2[p])$

25　end

26 end

27 对测试集中的每一个实例 t_i,先找到它的 K 个近邻,签为 $N(t_i)$,$i\in\{1,2,\cdots,n\}$

28 for　$i\in\{1,2,\cdots,q\}$ do

29　$Y1(i)=P1(i)*PEH1(i,d)$

30　$Y0(i)=P0(i)*PEH0(i,d)$ //d 为近邻中恰好包含标签 Li 的个数

31 end

32 for　$i\in\{1,2,\cdots,q(q-1)/2)\}$ do

33　$YY1(i)=\sum Q1_{ij}*PEH1(ij,d)$

34　$YY0(i)=\sum Q0_{ij}*PEH0(ij,d)$　// ij 表示由 L_i 和 L_j 组合后对应的新标记($i\neq j$)

35　$y1(i)=\alpha*Y1(i)+(1-\alpha)*YY1(i)$

36　$y0(i)=\alpha*Y0(i)+(1-\alpha)*YY0(i)$

37　if ($y1(i)>y0(i)$)　t 包含标签 Li

38　　else　t 不包含标签 Li

39 end

3　实验与分析

3.1　评价指标

实验采用 Subset Accuracy(Subacc)、Hamming Loss(hloss)、Accuracy(Acc)、Precision(Pre)、Recall(Rec) 这五个性能评价指标,具体计算如式(11) ~(15)所示。

$$\text{Subsetacc}(h)=\frac{1}{p}\sum_{i=1}^{p}[h(x_i)=Y_i] \tag{11}$$

其中: 对于 $[x]$,当 x 为真时取值为 1,反之为 0。该指标取值越大,则系统性能越优。

$$\text{hloss}(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|h(x_i)\Delta Y_i| \tag{12}$$

其中: 算子 Δ 用于度量两个集合之间的对称差,算子 $|.|$ 用于返回集合的势。该评价指标用于考察样本在单个标记上的误分类情况,取值越小,则系统性能越优。

$$\text{Accuracy}=\frac{TP_j+TN_j}{TP_j+FP_j+TN_j+FN_j} \tag{13}$$

$$\text{Precision}=\frac{TP_j}{TP_j+FP_j} \tag{14}$$

$$\text{Recall}=\frac{TP_j}{TP_j+FN_j} \tag{15}$$

其中: TP_j = 真正例个数,FP_j = 伪正例个数,TN_j = 真负例个数,FN_j = 伪负例个数。

3.2　实验设置

实验采用三个不同领域的多标签数据集(其中标签的势为样本的平均标签个数,标签密度为标签的势与标签总数的比值),分为训练数据和测试数据两大类,具体信息如表 2、3 所示。

表 2　训练数据集的相关信息

	样本数量	标签个数	标签的势	标签密度
Arts	2 000	26	1.627	0.063
Business	2 000	30	1.590	0.053
Computers	2 000	33	1.487	0.045

表 3　测试数据集的相关信息

	样本数量	标签个数	标签的势	标签密度
Arts	3 000	26	1.642	0.063
Business	3 000	30	1.586	0.053
Computers	3 000	33	1.522	0.046

3.3　实验结果及分析

实验结果从 3.1 节所述的几个评价指标对所提出的 S-ML-kNN 算法与传统的 ML-kNN 方法进行比较。具体的参数设置和实验结果如表 4 ~7 所示。

表 4　$\alpha=0.25$

指标	$K=3$		$K=5$		$K=10$	
	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN
Subacc	0.763 3	0.722 1	0.780 0	0.792 7	0.789 9	0.794 6
hloss	0.134 3	0.145 7	0.129 8	0.134 7	0.100 3	0.112 3
Acc	0.821 7	0.834 9	0.823 8	0.836 1	0.837 1	0.840 2
Pre	0.838 5	0.838 9	0.843 6	0.846 9	0.848 5	0.851 6
Rec	0.849 5	0.850 8	0.853 3	0.856 7	0.857 9	0.862 8

表 5　$\alpha=0.50$

指标	$K=3$		$K=5$		$K=10$	
	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN
Subacc	0.830 0	0.730 6	0.845 7	0.801 3	0.867 1	0.824 7
hloss	0.119 5	0.120 3	0.105 6	0.106 9	0.099 3	0.109 4
Acc	0.834 9	0.845 3	0.849 0	0.851 2	0.897 8	0.903 4
Pre	0.849 7	0.856 6	0.893 5	0.897 7	0.905 6	0.909 3
Rec	0.853 2	0.860 3	0.903 1	0.908 7	0.918 4	0.920 5

表 6　$\alpha=0.75$

指标	$K=3$		$K=5$		$K=10$	
	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN
Subacc	0.887 6	0.889 8	0.899 2	0.903 3	0.910 1	0.917 6
hloss	0.097 5	0.094 2	0.091 5	0.089 6	0089 9	0.083 4
Acc	0.872 1	0.874 5	0.883 3	0.886 7	0.897 8	0.903 4
Pre	0.901 6	0.907 8	0.925 9	0.926 5	0.946 5	0.948 9
Rec	0.913 5	0.917 8	0.943 4	0.958 3	0.950 2	0.953 4

表 7　$\alpha=0.90$

指标	$K=3$		$K=5$		$K=10$	
	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN	ML-kNN	S-ML-kNN
Subacc	0.878 2	0.894 3	0.886 2	0.897 8	0.894 2	0.890 8
hloss	0.103 2	0.101 2	0.095 1	0.092 3	0.092 5	0.090 2
Acc	0.884 9	0.885 9	0.893 1	0.896 9	0.895 5	0.890 3
Pre	0.897 2	0.879 8	0.903 4	0.906 5	0.907 9	0.914 6
Rec	0.904 6	0.909 7	0.916 2	0.918 1	0.913 6	0.922 7

表 4 ~7 给出取不同 α 值时 S-ML-kNN 算法和传统的 ML-kNN 算法在三个基准数据集上的对比实验结果,且每个表中近邻个数 K 的取值有 $K=3$、5、10 三种。

实验中各项指标均为 Arts、Business、Computers 三个数据集求解结果的平均值。实验结果表明,对于近邻的取值 K,一

般取值越大，各项指标显示的系统性能越好，这是因为随着 K 值的增大，对训练数据集中示例的近邻包含标签的情况可以获取更多信息，从而可以更准确地预测未知示例的标签包含情况，但是较大的 K 值需要找到更多的近邻，从而增加了算法计算的复杂度。而对于参数 α，有 $0 \leqslant \alpha \leqslant 1$。当 $\alpha = 1$，对应的 S-ML-kNN 算法即为传统的 ML-kNN 方法。在取值变化过程中，对于包含不同标签集的数据集而言，α 并不是取值越大越好，而是与示例个数及标签集大小密切相关，需要多次取值求最优解。上述实验结果得到的表 2 反映出 α 的最优值为 0.75。

表 8 给出了 S-ML-kNN 与其他算法的比较实验数据。结果表明，在这三个基准数据集上，S-ML-kNN 算法整体上表现优于其他三种算法。

表 8 RAkEL、CLR、C4.5、ML-kNN、S-ML-kNN 五种算法的对比($K = 10$，$\alpha = 0.75$)

算法	RAkEL	CLR	C4.5	ML-kNN	S-ML-kNN
Subacc	0.894 3	0.891 6	0.807 4	0.832 4	0.890 8
hloss	0.101 2	0.092 3	0.103 1	0.100 1	0.090 2
Acc	0.885 9	0.896 9	0.875 3	0.881 7	0.890 3
Preon	0.879 8	0.906 5	0.911 3	0.892 8	0.914 6
Rec	0.909 7	0.918 1	0.905 7	0.903 1	0.922 7

4 结束语

在多标签分类中，标签之间的相关性是一个不可忽略的重要因素。为了充分利用标签之间的相关性来改善多标签分类的性能，本文提出了一种联系标签相关性的 ML-kNN 算法，该算法使用标签间的共现概率大小来表示标签与标签之间相关性的强弱。实验结果表明，融合了标签之间相关性的 ML-kNN 算法在原来的 ML-kNN 算法的基础上可以改善多标签分类的性能。

参考文献:

[1] SCHAPIRE R E, SINGER Y. Boostexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2-3): 135-168.

[2] GODBOLE S, SARAWAGI S. Discriminative methods for multi-labele classification [C]//Proc of the 8th Pacic-Asia Conference on Knowledge Discovery and Data Mining. 2004: 22-30.

[3] 卫志华. 中文文本多标签分类研究 [D]. 上海: 同济大学, 2010.

[4] XU Xin-shun, JIANG Yuan, PENG Liang, *et al*. Ensemble approach based on conditional random field for multi-label image and video annotation [C]//Proc of the 19th ACM International Conference on Multimedia. 2011: 1377-1380

[5] QIN Jian-zhao, YUNG N H C. Feature fusion within local region using localized maximum-margin learning for scene categorization [J]. Pattern Recognition, 2012, 45(4): 1671-1683.

[6] GTZANIS, BERBERIDIS C, VLAHAVAS I. Machine learning and data mining in bioinformatics [M]//Handbook of Research on Innovations in Database Technologies and APPlications: Current and Future Trends. 2009.

[7] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Data mining and knowledge discovery handbook [M]. 2nd ed. [S. l.]: Springer, 2010: 667-685.

[8] TROHIDIS K. Multi-label classification of stack binary relevance models for mutilabel classifiers [J]. Eurasip Journal on Audio Speech and Music Processing, 2011, 4(1): 4-6.

[9] FURNKRANZ J, HULLERMEIER E, MENCIA E L, *et al*. Multi-label classification via calibrated label ranking [J]. Machine Learning, 2008, 73(2): 133-152.

[10] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multilabel classification [C]//Proc of the 18th European Conference on Machine Learning. 2007: 406-417.

[11] ZHANG Min-ling, ZHOU Zhi-hua. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Trans on Knowledge and Data ENgineering, 2006, 18(10): 1338-1351.

[12] ZHANG Min-ling, ZHOU Zhi-hua. ML-kNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[13] ZHANG Min-ling, ZHANG Kun. Multi-label learning by exploiting label dependency [C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC: IEEE Computer Society, 2010: 999-1008.

[14] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval [M]. New York: ACM Press, 1999: 123-129.

(上接第 2608 页) 更好的优势。下一步的工作是进一步优化离群点排除机制，提高不确定数据流聚类质量。

参考文献:

[1] NGAI W K, KAO Ben, CHUI C K, *et al*. Efficient clustering of uncertain data [C]//Proc of the 6th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2006: 436-445.

[2] KRIEGEL H P, PFEIFLE M. Density-based clustering of uncertain data [C]//Proc of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2005: 672-677.

[3] KRIEGEL H P, PFEIFLE M. Hierarchical density-based clustering of uncertain data [C]//Proc of the 5th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2005: 689-692.

[4] CALLAGHAN L O, MISHRA N, MEYERSON A, *et al*. Streaming-data algorithms for high-quality clustering [C]//Proc of the 18th International Conference on Data Engineering. San Jose: IEEE Press, 2002: 685-694.

[5] ZHU Wei-heng, YIN Jian, XIE Yi-huang. Arbitrary shape cluster algorithm for clustering data stream [J]. Journal of Software, 2006, 17(3): 379-387.

[6] AGGARWAL C C, HAN Jia-wei, WANG Jian-yong, *et al*. A framework for clustering evolving data streams [C]//Proc of the 29th International Conference on Very Large Data Bases. Berlin: Morgan Kaufmann Publishers, 2003: 81-92.

[7] ZHANG Chen, GAO Ming, ZHOU Ao-ying. Tracking high quality clusters over uncertain data streams [C]//Proc of the 25th IEEE International Conference on Data Engineering. [S. l.]: IEEE Press, 2009: 1641-1648.

[8] 张晨，金澈清，周傲英. 一种不确定数据流聚类算法 [J]. 软件学报, 2010, 21(9): 2173-2187.

[9] 罗清华，彭宇，彭喜元. 一种多维不确定性数据流聚类算法 [J]. 仪器仪表学报, 2013, 34(6): 1330-1338.

[10] CAO Feng, ESTERY M, QIAN Wei-ning, *et al*. Density-based clustering over an evolving data stream with noise [C]//Proc of the 6th SIAM International Conference on Data Mining. Bethesda: SIAM, 2006: 326-337.

[11] 颜一鸣，郭鑫. 一种新的不确定树模式聚类算法 [J]. 计算机工程与科学, 2013, 35(7): 156-163.

[12] RÉ C, LETCHNER J, BALAZINSKA M, *et al*. Event queries on correlated probabilistic streams [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008: 715-728.

[13] 胡春安，范丽文，毛伊敏. HPDBSCAN: 高效的不确定数据处理算法 [J]. 计算机工程与设计, 2013, 34(3): 1044-1049.

[14] 郭鑫，颜一鸣，徐洪智，等. 不确定树数据库中的动态聚类算法 [J]. 小型微型计算机系统, 2013, 34(6): 1339-1343.