基于概念权重向量的MIMLSVM改进算法

环 天 郝 宁 牛 强
(中国矿业大学计算机科学与技术学院 徐州 221116)

摘 要 针对多示例多标记学习算法MIMLSVM只从包层面构造聚类,而忽略了包内示例分布对分类造成影响这一不足,提出一种基于概念权重向量的MIMLSVM改进算法——I-MIMLSVM算法。首先从示例层面构造聚类,挖掘出示例中的潜在概念簇,运用R-PATTERN算法计算每个概念簇的概念权重;然后利用TF-IDF算法计算每个概念簇在各个示例包中的重要度;最后将示例包表示为概念权重向量,向量的每一维即为概念簇的概念权重与其在该包中的重要度的乘积。将该算法在包含2000幅图像的自然数据集上进行实验验证,结果表明改进的算法在分类性能上整体优于原算法,尤其在Hamming loss,Coverage和Average precision这3个测评指标上较为明显。

关键词 MIMLSVM,聚类,R-PATTERN,TF-IDF
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.009

Improved MIMLSVM Algorithm Based on Concept Weight Vector

HUAN Tian HAO Ning NIU Qiang
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract In order to solve the problem that the MIMLSVM algorithm only constructs cluster from the bag level, while ignoring the distribution of the instance in the bag, this article proposed an improved MIMLSVM algorithm I-MIMLSVM. Firstly, we constructed clustering from the sample level, and explored the potential cluster of concepts in the example. Then, we used R-PATTERN algorithm to calculate the weight of each concept cluster, and calculated the importance degree of each concept cluster in each bag with TF-IDF algorithm. Finally, each bag was represented as a concept vector, and each dimension of the vector was equal to the multiplication of the weight of each concept cluster and its importance degree in this bag. The natural data set containing 2000 images was used in our experiments. The experimental results show that the improved algorithm performs better than the original algorithm, especially in the Harming loss, Coverage and Average precision.

Keywords MIMLSVM, Cluster, R-PATTERN, TF-IDF

1 引言

近年来,随着人工智能领域的飞速发展,机器学习[1]迅速成为该领域的一大研究热点,监督学习[2]又是其中研究最多、应用最广泛的一种学习框架[3]。但是传统监督学习的局限在于无法有效处理歧义性对象,造成这种局限的主要原因是其采用单示例单标记的形式描述现实对象,无法准确完整地刻画出对象的真实信息,因此需要一种新型的学习框架来处理歧义性对象。

多示例多标记(MIML)学习[4]是基于对象的歧义性而提出的一种学习框架,是对多示例学习[5]和多标记学习[6]的扩展[7]。MIML学习采用多个示例和多个标记的形式描述对象,它虽然相对于单示例单标记更加复杂,但符合现实对象的语义特征,有利于学习任务的完成和实际运用。

目前,已经出现了多种MIML学习算法,如具有高效率的MIMLfast算法[8]、基于正则化和最大化间隔的D-MIMLSVM[9]以及M^{3}MIML算法、基于多视图的MIMLmix[10]算法、基于多模态图像标注的M3LDA[11]算法等。其中MIMLSVM是一种基于多标记学习退化策略的算法[12],该算法先将示例包集合划分为若干个聚类,然后用示例包和各个聚类中心的距离重新表示每个示例包,从而多示例多标记样本就可以转化为单示例多标记的样本;最后利用文献[13]中的MLSVM算法对转化后的样本进行学习求解。实验证明,该算法在场景分类和文档分类问题中取得了良好的分类效果。

2 MIMLSVM算法分析

给定MIML样本(X_i,Y_i),设$z_i=\Phi(X_i)$,$\Phi:2^X\rightarrow Z$,$\varphi:Z\times Y\rightarrow(+1,-1)$,对$\forall y\in Y$,有$\varphi(z_i,y)=+1$当且仅当$y\in Y_i$,否则$\varphi(z_i,y)=-1$。首先,MIML样本$(X_u,Y_u)$中的示例

到稿日期:2016-10-11 返修日期:2016-11-12 本文受江苏省产学研联合创新资金前瞻性联合研究项目(BY2014028-09)资助。
环 天(1992-),女,硕士生,主要研究方向为传感器定位、人工智能,E-mail:Htian2010@cumt.edu.cn;郝 宁(1990-),男,硕士生,主要研究方向为人工智能、智能信息处理;牛 强(1974-),男,博士,教授,主要研究方向为数据挖掘、智能优化算法、智能信息处理、机器学习。

包 X_u 被统一归入到集合 Λ 中。然后,对集合 Λ 运用 k-medoids 聚类算法,这里利用 Hausdorff[14] 刻画集合间的距离。给定示例包 $A=\{a_1,a_2,\cdots,a_{nA}\}$,$B=\{b_1,b_2,\cdots,b_{nB}\}$,则 A 和 B 的 Hausdorff 距离可定义为:

$$d_H(A,B)=\max\{\max_{a\in A}\min_{b\in B}\|a-b\|,\max_{b\in B}\min_{a\in A}\|b-a\|\} \tag{1}$$

其中,$\|a-b\|$ 表示示例 a 和 b 间的 Euclidean 距离[15]。

聚类之后,集合 Λ 被分为 k 个类,聚类中心为 $M_t(t=1,2,\cdots,k)$,基于构造性聚类将示例包 X_u 转化为一个 k 维的数值向量 z_u,z_u 的每一维即为 X_u 到各个聚类中心的距离。这样,原 MIML 样本 $(X_u,Y_u)(u=1,2,\cdots,m)$ 即可转化为单示例多标记样本 (z_u,Y_u)。最后,利用 MLSVM 算法学习得到 f_{MLL},则 $f_{MIML}(X^*)=f_{MLL}(z^*)$。虽然以上转化会丢失部分信息,但 MIML 相对于其他框架可以获得比对象更多的原始信息,因此 MIMLSVM 仍会取得良好的学习效果。假设 X^* 为给定的新样本,则 MIMLSVM 算法的具体过程如下:

(1)对于 MIML 样本 $(X_u,Y_u)(u=1,2,\cdots,m)$,设集合 $\Lambda=\{X_u\mid u=1,2,\cdots,m\}$;

(2)从 Λ 中随机选择 k 个元素作为初始聚类中心 $M_t(t=1,2,\cdots,k)$,重复以下步骤,直到所有的 M_t 不再改变:

1)$\Lambda_t=\{M_t\}(t=1,2,\cdots,k)$;

2)对于每个 $X_u\in(\Lambda-\{M_t\mid t=1,2,\cdots,k\})$,分别计算 $index=\arg\min_{t\in\{1,2,\cdots,k\}}d_H(X_u,M_t)$,$\Lambda_{index}=\Lambda_{index}\cup\{X_u\}$;

3)$M_t=\arg\min_{A\in\Lambda_t}\sum_{B\in\Lambda_t}d_H(A,B)(t=1,2,\cdots,k)$;

(3)将 (X_u,Y_u) 转化为多标记样本 $(z_u,Y_u)(u=1,2,\cdots,m)$,其中 $z_u=(z_{u1},z_{u2},\cdots,z_{uk})=(d_H(X_u,M_1),d_H(X_u,M_2),\cdots,d_H(X_u,M_k))$;

(4)对于 $y\in Y$,可得到数据集 $D_y=\{(z_u,\varphi(z_u,y))\mid u=1,2,\cdots,m\}$,然后训练得到 SVM[16] 分类器 $h_y=SVMTrain(D_y)$;

(5)得到 $Y^*=\{\arg\max_{y\in Y}h_y(z^*)\}\cup\{y\mid h_y(z^*)\geqslant 0,y\in Y\}$,其中 $z^*=(d_H(X^*,M_1),d_H(X^*,M_2),\cdots,d_H(X^*,M_k))$。

3　算法分析与改进

虽然 MIMLSVM 算法取得了良好的实验效果,但其转化过程涉及到的构造性聚类方法原先是从示例层聚类将多示例样本转化为单示例样本,而这里却从包层面对 MIML 样本进行转化,没有考虑示例包的内在结构,忽略了示例分布对包标记的潜在影响。

本文结合文本处理中的 R-PATTERN 算法[17] 以及 TF-IDF 算法[18],提出了一种新的示例包表示方法。首先,通过对示例空间中的所有示例进行聚类,得到能够反映示例潜在概念的簇;然后,利用 R-PATTERN 计算每个概念簇的权重,再利用 TF-IDF 算法计算每一个概念簇对于每个包的重要度;最后,将每个示例包转化成单个示例,此时,示例的每一维即为各个概念簇对该包的重要度与该概念权重的乘积。此种表示方法从示例层次出发,不仅能够有效度量示例对包标记的影响,还能够评估每个概念簇对整个示例空间的重要度。

3.1　概念权重

本文借助在文本分类中被成功应用的 R-Pattern 方法评估示例中潜藏概念的权重。R-Pattern 是一种基于关键词统计的文本向量表达技术,即将每篇文件的关键词向量看作一个模式,若两个文件中含有相同的关键词,则合并两个文件的模式,合并后的模式称为 R-Pattern。

为了挖掘出潜藏在示例中的概念,本文选用已在原 MIMLSVM 算法中得到应用的 k-medoids 聚类算法,将训练集 D 中的所有示例划分为 k 个概念簇,即 $C=\{C_1,C_2,\cdots,C_k\}$,把示例包和概念类比为文件和关键词,则可以计算每个示例包相对于所有概念簇的 R-Pattern。

定义 1　对于概念簇 $C=\{C_1,C_2,\cdots,C_k\}$,定义 $count(B,c)$ 为计算每个多示例包 B 中所包含属于概念簇 c 的示例数目,则有:

$$p_i=\{(c,f)\mid c=C,f=count(B,c)>0\} \tag{2}$$

若两个 R-Pattern 所包含的概念簇一样,则可以进行合并,表 1 和表 2 列出了合并过程。

表 1　示例包的概念构成

包	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Bag1	2	1	0	0	1	1	0
Bag2	0	0	2	1	0	1	2
Bag3	0	0	3	1	1	1	0
Bag4	0	0	1	1	1	1	0
Bag5	1	1	0	0	0	1	1
Bag6	2	1	0	0	0	1	1

表 2　示例包的 R-Pattern

R-Pattern	C_1	C_2	C_3	C_4	C_5	C_6	C_7	*Sup*
Rp1	2	1	0	0	1	1	0	1/6
Rp2	0	0	2	1	0	1	2	1/6
Rp3	0	0	4	2	2	2	0	1/3
Rp4	3	2	0	0	0	2	2	1/3

计算得到每个概念簇对于各个 R-Pattern 的重要度 $WP=[w_{ij}]_{r\times k}$,如表 3 所列。其中,Rp 表示 R-Pattern,Sup 表示每个 R-Pattern 的支持度,其值为 R-Pattern 中包含的示例包个数与所有示例包个数的比值,$w_{ij}=Rp_{ij}/n_i$,Rp_{ij} 为第 i 个 R-Pattern 的第 j 个域,n_i 为第 i 个 R-Pattern 包含的示例包数目,r 为 R-Pattern 的总数,k 为所有的概念簇数目。以表 3 为例,Rp1={Bag1},Rp2={Bag2},Rp3={Bag3,Bag4},Rp4={Bag5,Bag6},则可得每个概念的重要度。

表 3　R-Pattern 的概念重要度

R-Pattern	C_1	C_2	C_3	C_4	C_5	C_6	C_7	*Sup*
Rp1	2/1	1/1	0	0	1/1	1/1	0	1/6
Rp2	0	0	2/1	1/1	0	1/1	2/1	1/6
Rp3	0	0	4/2	2/2	2/2	2/2	0	1/3
Rp4	3/2	2/2	0	0	0	2/2	2/2	1/3

定义 2　对于概念簇 $C=\{C_1,C_2,\cdots,C_k\}$,$WP=[w_{ij}]_{r\times k}$ 为每个概念簇对于各个 R-Pattern 的重要度,则概念权重为:

$$WC_i=\sum_{j=1}^{r}w_{ij}/sup_j \tag{3}$$

其中,r 为 R-Pattern 的数目,sup_j 为第 j 个 R-Pattern 的支持度。

3.2　概念权重向量

TF-IDF 算法是一种用于资讯检索与资讯探勘的常用加权技术,可以评估一个字词对于一个文件集或一个语料库中

一份文件的重要程度。其主要思想是：若某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适用于分类。

TF-IDF 包含两个部分：TF（词频）和 IDF（逆向文件频率）。TF 指某一个给定的词语在某个文件中出现的次数；IDF 指一个词语普遍重要性的度量。对于某一文件中的词语 t_i，其 TF 可表示为：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

IDF 可表示为：

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{5}$$

式(4)、式(5)中，$n_{i,j}$ 表示 t_i 在文件 d_j 中出现的次数，$\sum_k n_{k,j}$ 表示 d_j 中所有字词出现的次数之和，$|D|$ 指资料库中的文件总数，$|\{j:t_i\in d_j\}|$ 指包含词语 t_i 的文件数目，两者的乘积即为词语 t_i 的 TF-IDF 值。根据某一特定文件内的高词语频率以及该词语在整个文件集合中的低文件频率，可以得到高权重的 TF-IDF。因此，TF-ID 算法可以保留重要的词语。

通过将示例包和概念类比作文件和关键词，利用 R-Pattern 方法计算得到了示例中潜在概念的概念权重。结合表 1 列出的示例包概念构成，同样可以利用 TF-IDF 算法计算每个概念簇在每个示例包中的重要度，于是有：

$$W_{TF\text{-}IDF(i,j)}=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{|\{j:C_i\in Bag_j\}|+1} \tag{6}$$

其中，$n_{i,j}$ 表示概念簇 C_i 包含属于示例包 Bag_j 的示例的个数，$\sum_k n_{k,j}$ 表示概念簇 C_i 含有的示例个数，$|D|$ 表示训练集中示例包的总数，$|\{j:C_i\in Bag_j\}|$ 表示概念簇 C_i 涵盖示例包的个数。

结合计算得到的概念权重，MIML 样本中的多示例包可以表示为如下的 k 维向量，这里称其为概念权重向量。

$$Bag_j=[(W_{TF\text{-}IDF(1,j)}\times WC_1),(W_{TF\text{-}IDF(2,j)}\times WC_2),\cdots,W_{TF\text{-}IDF(k,j)}\times WC_k] \tag{7}$$

向量的每一维为由 TF-IDF 计算得到的第 i 个概念簇在 Bag_j 中的重要度与 R-Pattern 计算得到的第 i 个概念簇的概念权重的乘积。

3.3 算法描述

根据上文思想，针对多示例多标记学习算法 MIMLSVM 只从包层面构造聚类，而忽略包内示例分布对分类造成影响这一不足，提出一种基于概念权重向量的 MIMLSVM 改进算法——I-MIMLSVM 算法。首先从示例层面构造聚类，挖掘出示例中的潜在概念簇，运用 R-PATTERN 算法计算每个概念簇的概念权重；然后利用 TF-IDF 算法计算每个概念簇在各个示例包中的重要度；最后将示例包表示为概念权重向量，向量的每一维即为概念簇的概念权重与在该包中重要度的乘积。算法具体描述如算法 1 所示。

算法 1 I-MIMLSVM 算法

输入：训练示例空间 X，标记空间 Y，聚类簇数 k，聚类算法 k-medoids，R-Pattern 算法，MLSVM 算法

输出：MIML 分类器

1. 利用聚类算法 k-medoids 将示例空间 X 划分为 k 个概念簇 $C=\{C_1,C_2,\cdots,C_k\}$；
2. 将概念簇和示例包类比为关键词和文件，将示例包中的示例量化到 k 个概念簇上，如表 1 所列，得到示例包的概念构成；
3. 合并具有相同概念簇的示例包，构成向量，得到示例包的 R-Pattern，如表 2 所列；
4. 计算每个概念簇对于各个 R-Pattern 的重要度 $WP=[w_{ij}]_{r\times k}$，然后可得到每个概念簇的概念权重 $WC_i=\sum_{j=1}^{r} w_{ij}/sup_j$；
5. 再利用 TF-IDF 计算每个概念簇对于各个示例包的重要度 $W_{TF\text{-}IDF(i,j)}$；
6. 将每个示例包表示成概念权重向量 $Bag_j=[(W_{TF\text{-}IDF(1,j)}\times WC_1),(W_{TF\text{-}IDF(2,j)}\times WC_2),\cdots,W_{TF\text{-}IDF(k,j)}\times WC_k]$；
7. 最后利用 MLSVM 算法对转化过后的单示例多标记训练集求解得到分类器。

4 实验与分析

4.1 实验环境

选用自然场景分类数据作为实验数据集，该数据集包含 2000 幅自然场景图像，所有可能的概念类为 desert，mountains，sea，sunset 以及 trees，数据集中具有两个或两个以上标记的图像约占数据集的 22%。表 4 列出了该数据集的具体描述信息。

表 4 自然场景数据详情

Label	Images	Label	Images
Desert	340	mountains+sunset	19
Mountains	268	mountains+trees	106
Sea	341	sea+sunset	172
Sunset	216	sea+trees	14
Trees	378	sunset+trees	28
desert+mountains	19	desert+mountains+sunset	1
desert+sea	5	desert+sunset+trees	3
desert+sunset	21	mountains+sea+trees	6
desert+trees	20	mountains+sunset+trees	1
mountains+sea	38	sea+sunset+trees	4

4.2 实验结果分析

算法的性能评估选用以下评测指标：Hamming loss ↓，one-error ↓，Coverage ↓，ranking loss ↓，Average precision ↑（↓表示值越小越好，↑表示值越大越好）。本文选用 k-medoids 算法作为聚类算法，并与 MIMLSVM 算法和 MLSVM 算法进行比较，实验结果如表 5－表 7 所列。

表 5 基于概念权重向量的改进算法(I-MIMLSVM)的实验结果

高斯内核	评测指标				
	Hamming loss	one-error	Coverage	ranking loss	Average precision
$\gamma=0.2$	0.178±0.018	0.330±0.031	1.019±0.083	0.189±0.019	0.786±0.022
$\gamma=0.3$	0.185±0.018	0.345±0.033	1.063±0.095	0.199±0.018	0.776±0.020
$\gamma=0.4$	0.191±0.013	0.360±0.028	1.077±0.097	0.204±0.022	0.768±0.023

表 6　MIMLSVM 算法的实验结果

高斯内核	评测指标				
	Hamming loss	one-error	Coverage	ranking loss	Average precision
$\gamma=0.2$	0.180±0.017	0.327±0.033	1.022±0.085	0.187±0.018	0.783±0.020
$\gamma=0.3$	0.188±0.016	0.344±0.032	1.065±0.094	0.196±0.020	0.772±0.020
$\gamma=0.4$	0.193±0.014	0.358±0.030	1.080±0.099	0.202±0.022	0.764±0.021

表 7　MLSVM 算法的实验结果

高斯内核	评测指标				
	Hamming loss	one-error	Coverage	ranking loss	Average precision
$\gamma=0.2$	0.196±0.013	0.368±0.032	1.115±0.122	0.211±0.023	0.756±0.022
$\gamma=0.3$	0.195±0.015	0.370±0.034	1.129±0.113	0.214±0.022	0.754±0.023
$\gamma=0.4$	0.196±0.016	0.372±0.034	1.151±0.122	0.220±0.024	0.751±0.023

将表 5 和表 6 分别与表 7 作对比,I-MIMLSVM 算法和 MIMLSVM 算法在每个评测指标上均优于 MLSVM 算法,且数据差距较明显。表 5 和表 6 的数据较接近,不便观察比较,因此给出了表 5 和表 6 的柱状图,为了便于区别,将两表中的 Coverage 和 Average precision 的值分别减去 1 和 0.7。处理结果如图 1－图 5 所示,可以看出 MIMLSVM 算法和 I-MIMLSVM算法均在 $\gamma=0.2$ 时取得较好的实验结果。比较图 1、图 3 和图 5,改进算法在 Hamming loss,Coverage,Average precision 3 个指标上取得了比 MIMLSVM 算法更好的实验效果,在其他两个评测指标上则略差于 MIMLSVM 算法。由图 2 和图 4 可以看出,两者差距不大,实验结果基本相当。实验表明,本文提出的基于概念权重向量表示多示例包的方法是有效的,学习效果在整体上略优于原算法。

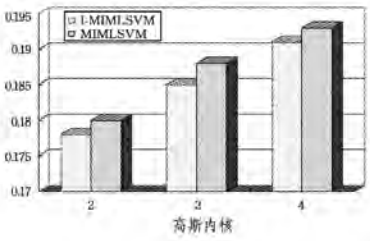

图 1　Hamming loss 对比图

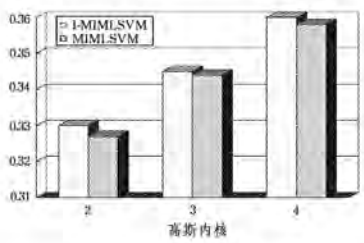

图 2　one-error 对比图

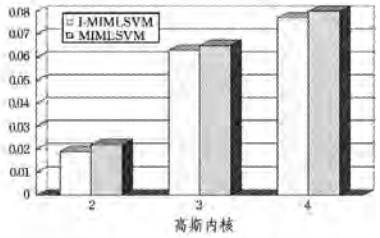

图 3　Coverage 对比图

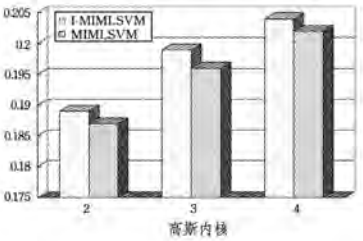

图 4　ranking loss 对比图

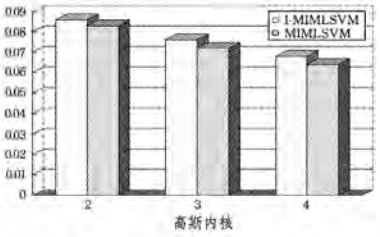

图 5　Average precision 对比图

结束语　本文主要研究了基于多标记学习退化策略的 MIMLSVM 算法,并针对算法的不足之处作出改进。在分析了 MIMLSVM 算法的基本原理后,针对算法转化过程中忽视了示例对包标记的影响的问题,提出了概念权重向量的表示形式的解决方法,弥补了 MIMLSVM 算法的不足之处。实验结果表明,改进后的算法能有效提高分类性能,与原算法相比具有优势。

参 考 文 献

[1] NATARAJAN B B K. Machine Learning: A Theoretical Approach [J]. IEEE Expert Intelligent Systems & Their Applications, 2015, 7(4): 89-90.

[2] COHEN M E, HUDSON D L. Neural Networks and Artificial Intelligence for Biomedical Engineering [M]. Wiley-IEEE Press, 1999: 10676-10681.

[3] MURPHY K P. Machine Learning: A Probabilistic Perspective [J]. Mathematics Education Library, 2012, 58(8): 27-71.

[4] ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification [C] // The 2006 Conference Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 1609-1616.

[5] ZHANG D, HE J R, LAWRENCE R. MI2LS: multi-instance learning from multiple information sources [C] // The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 149-157.

(下转第 63 页)

式的聚类算法 PatDistPCM 和 PatCluPCM 可以有效减小数据的偏离性对聚类结果的影响，提高聚类的准确程度。而在准确性指标几乎相同的情况下，PatCluPCM 算法的效率远高于 PatDistPCM 算法，因此在数据量较大的情况下 PatClu-PCM 算法更适合于非随机缺失数据的聚类。

结束语　数据缺失模式是数据缺失机制的直接反映。本文提出的两种模糊聚类算法 PatDistPCM 和 PatCluPCM 分别通过最小化簇内缺失模式距离之和以及对缺失模式聚类两种方式，利用缺失模式中隐藏的缺失机制信息对不完整数据进行聚类，当数据缺失属于非随机缺失时这部分信息对聚类的结果会产生重大影响。实验结果表明，通过引入缺失模式能够提高 PCM 算法对非随机缺失数据的聚类精度。

参 考 文 献

[1] HAN J W, KAMBER M, PEI J. Data Mining: Concepts and Techniques(3rd ed)[M]. Morgan Kaufmann Publishers, 2011: 288-293.

[2] GU Y, YU G, LI X J, et al. RFID data interpolation algorithm based on dynamic probabilistic path-event model[J]. Journal of Software, 2010, 21(3): 438-451.

[3] DIXON J K. Pattern recognition with partly missing data[J]. IEEE Transactions on Systems, Man and Cybernetics, 1979, 9(10): 617-621.

[4] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. Plenum Press, 1981.

[5] HATHAWAY R J, BEZDEK J C. Fuzzy c-Means Clustering of Incomplete Data[J]. IEEE Transactions on System, Man, and Cybernetics, 2001, 31(5): 735-744.

[6] BALKIS A, YAHIA S B. A new algorithm for fuzzy clustering handling incomplete dataset[J]. International Journal on Artificial Intelligence Tools, 2014, 23(4): 1460012.

[7] KRISHNAPURAM R, KELLER J M. A possibilistic Approach to clustering[J]. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98-110.

[8] ZHANG Q, CHEN Z. A distributed weighted Possibilistic c-Means algorithm for clustering incomplete big sensor data[J]. International Journal of Distributed Sensor Networks, 2014, 2014(2): 4.

[9] LITEEL R J A, RUBIN D B. Statistical Analysis with Missing Data[M]. John Wiley & Sons, Inc. New Jersey, 2002.

[10] Donald D B. Inference and Missing Data[J]. Biometrika, 1976, 63(3): 581-592.

[11] ALLISON P D. 数据缺失[M]. 林毓玲，译. 上海：格致出版社，2012.

[12] MARLIN B M. Missing Data Problems in Machine Learning [D]. Toronto: University of Toronto, 2008.

[13] MARLIN B M, ZEMEL R S. Collaborative Prediction and Ranking with Non-Random Missing Data [C] // RecSys '09. New York, USA, 2009: 23-25.

[14] WANG H, WANG S. Discovering patterns of missing data in servey databases: An application of rough sets[J]. Expert Systems with Applications, 2009, 36(3): 6256-6260.

[15] TIMM H, BORGELT C, KRUSE R. An Extension of Possibilistic Fuzzy Cluster Analysis[J]. Fuzzy Sets and Systems, 2004, 147(1): 3-16.

[16] BAGGA A, BALDWIN B. Entity-based cross-document coreferencing using the vector space model[C] // Proc. 1998 Annual Meeting of the Association for Computational Linguistics and Int. Conf. Computational Linguistics (COLING-ACL '98). Montreal, Quebec, Canada, 1998.

(上接第 51 页)

[6] ZHOU Z H, ZHANG M L. A Review on Multi-Label Learning Algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[7] GIBAJA E. A Tutorial on Multi-Label Learning [J]. ACM Computing Surveys, 2015, 47(3): 1-38.

[8] ZHANG M L, ZHOU Z H. M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning [J]. Eighth IEEE International Conference on Data Mining, 2008, 176(1): 688-697.

[9] ZHOU Z H, ZHANG M L, HUANG S J. Multi-instance Multi-label Learning [J]. Artificial Intelligence, 2008, 176(1): 2291-2320.

[10] NGUYEN C T, WANG X L, LIU J. Labeling complicated objects: Multi-view multi-instance multi-label learning[C] // The 28th AAAI Conference on Artificial Intelligence Quebec City. Canada: AAAI Press, 2014: 2013-2019.

[11] NGUYEN C T, ZHAN D C, ZHOU Z H. Multi-modal image annotation with Multi-instance Multi-label LDA[C] // International Joint Conference on Artificial Intelligence. Beijing: IJCAI, 2013: 1558-1564.

[12] ZHOU Z H, ZHANG M L, HUANG S J, et al. MIML: A Framework for Learning with Ambiguous Objects [J]. Corr Abs, 2008, 3231(1): 2012.

[13] BOUTELL M R, LUO J B, SHEN X P. Learning Multi-label Scene Classification [J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[14] KNAUER C, LÖFFLER M, SCHERFENBERG M. The directed Hausdorff distance between imprecise point sets [J]. Theoretical Computer Science, 2011, 412(32): 4173-4186.

[15] LIBERTI L, LAVOR C, MACULAN N. Euclidean distance geometry and applications [J]. Siam Review, 2012, 56(1): 3-69.

[16] KIM S, CHOI J. An SVM-based high-quality article classifier for systematic reviews [J]. Journal of Biomedical Informatics, 2014, 47(2): 153-159.

[17] ŽALIK K R, ŽALIK B. Validity Index for Clusters of Different Sizes and Densities [J]. Pattern Recognition Letters, 2011, 32(2): 221-234.

[18] HONG T P, LIN C W, YANG K T, et al. Using TF-IDF to Hide Sensitive Itemsets [J]. Applied Intelligence, 2013, 38(4): 502-510.