文章编号: 1001-0920(2009)04-0574-05

一种基于关联规则的多类标分类算法

李 宏, 李 博, 吴 敏, 陈松乔
(中南大学 信息科学与工程学院, 长沙 410083)

摘 要: 提出了一种基于关联规则的多类标算法(MLAC),利用多类标 FP-tree 来分解组合生成多类标规则,并通过组合多重关联规则分类器进行分类预测,降低了由高维属性带来的高计算复杂度,有效地提高了算法的性能和效率. 针对多类标数据集的实验结果表明,MLAC 算法在性能和效率等方面均优于 ML-KNN 等多类标分类算法.

关键词: 分类; 关联规则; 多类标; 单类标; 组合
中图分类号: TP301.6 **文献标识码**: A

Multi-label classification algorithm based on association rules

LI Hong, LI Bo, WU Min, CHEN Song-qiao
(School of Information Science and Engineering, Central South University, Changsha 410083, China. Correspondent: LI Hong, E-mail: lihongcsu@mail.csu.edu.cn)

Abstract: A multi-label classification algorithm based on association rules (MLAC) is proposed, which uses multi-label FP_tree to resolve multi-label data and merger labels with the same attributes to generate multi-label rules. The classification prediction is conducted by assembling the classifiers. So the computational complexity brought by high dimensional attribute decreases while the performance and efficiency increases. The results of subsequent experiment based on multi-label data show that MLAC achieves better performance and efficiency than ML-KNN and other multi-label classification algorithms.

Key words: Classification; Association rules; Multi-label; Single-label; Ensemble

1 引 言

从 Agrawal[1] 提出数据库中的关联规则挖掘后,关联规则挖掘算法及其应用得到迅速发展,关联规则的功能不再局限于概念描述. 1998 年,Liu 等[2] 提出了经典的基于分类关联规则(CAR)的关联分类算法:CBA (Classification based on association)算法.

关联分类技术一般包括两个步骤:1) 发现所有规则后件为类标签的分类关联规则(CARs);2) 从已发现的 CARs 中选择优先级高的规则来覆盖训练集. 规则的优先级往往根据分类关联规则的置信度、支持度、规则长度或一般分类规则质量标准进行评价[3]. 因为关联分类算法具有可解释性好、分类准确率高的诸多优点,在随后几年时间里,出现了很多新的关联分类算法,其中最有代表性的是 Li 等[4] 提出的基于多个分类关联规则的分类算法 CMAR (Accurate and efficient classification based on multiple class-association rules) 和 Yin[5] 提出的预测型关联规则的分类算法 CPAR.

这些传统的关联分类算法都只能对单类标数据进行处理,而在现实生活中,多类标分类的应用是很广泛的. 因为一个真实世界的对象很可能会和多个类别相关联. 例如,一个图片可以同时属于多个不同的类别,如关于雪山的图片就同时属于雪和山这两类. 目前,国内外针对多类标问题的研究还较少. Comité 等[6] 对二叉决策树进行扩展,用于处理多类标数据. 李宏等[7] 提出了一种有效的多类标决策树算法 SCC_SP,采用一种综合考虑集合同一性和一致性特征的类标集相似度的评定方法,来计算多类标数据的类标集相似度,进而作为属性分类效果的评定指标. Elisseeff 等[8] 采用支持向量机方法来处理多类标数据的分类问题,并对 Yeast 基因数据集进行测试,取得了较好的分类效果. 在基于多类标分类的组合算法中,Adaboost.MH[9] (A multi-class

收稿日期: 2008-02-26; **修回日期**: 2008-06-21.
基金项目: 国家杰出青年科学基金项目(60425310); 中南大学博士后基金项目(2008).
作者简介: 李宏(1966—),男,长沙人,教授,博士后,从事数据挖掘、信号处理的研究; 吴敏(1963—),男,广东化州人,教授,博士,从事鲁棒控制、过程控制等研究.

multi-lable version of adaboost based on hamming loss)是一种较为经典的算法,它引入了组合方法,从而提高了决策树算法的性能和稳定性. 周志华等[10] 提出 ML-KNN (Multi-label *K*-nearest neighbor)多类标学习算法,对于测试集中的每条样本,在训练集中寻找其 K 个最近邻样本,然后对这 K 个最近邻样本的类标集进行统计分析,计算这些近邻样本属于各个类标的概率,利用最大后验概率方法来确定当前未知样本的类标集. 该算法通过计算样本间距离来确定最近邻样本,没有建立模型. 所以,如何使用一种解释性好的算法来进一步提高预测精度是值得研究的问题.

基于关联规则的分类算法具有可解释性好、分类准确率较高的特点,但很少用于多类标学习的关联分类算法. Fadi 等[11]曾提出了基于关联规则的多类标学习算法 MMAC(A new multi-class, multi-label associative classification approach). 但因为其论文中并没有使用多类标数据集进行分类实验,无法验证 MMAC 算法针对多类标学习的效果. 本文通过分析关联分类算法进行多类标学习的难点,提出了一种基于多类标学习的关联分类算法 MLAC (Multi-label classification algorithm based on association rules),并将其与 ML-KNN, Adaboost.MH, MIMLSVM (Multi-instance and multi-label support vector machine), Rank-Svm(Rank support vector machine)等常见的多类标学习算法进行了实验比较.

2　相关理论和定义

2.1　关联规则相关定义

基于关联规则的分类是针对数据集挖掘频繁模式、建立分类器,实现对未知样本分类的方法. 给出相关定义如下:给定数据集 D, I 为 D 中所有项的集合, Y 为样本类别, R 为分类关联规则, X 为项的集合(项集), S 为 R 的支持数,sup 为 R 的支持度, $|D|$ 为数据集 D 的样本总个数,conf 为 R 的置信度, $N(X)$ 为 D 中含 X 的样本个数.

定义 1　R 具有如下形式: $R: X \Rightarrow Y$.

定义 2　R 的支持数 S 等于 D 中包含 X 且类别为 Y 的样本个数.

定义 3　一个规则 R 在 D 中实际出现次数 $N(r)$ 是 D 中与 R 的匹配的样本数.

定义 4　R 的支持度 $\sup(R) = S/|D|$.

定义 5　R 的置信度 $\mathrm{conf}(R) = S/N(X)$.

2.2　多类标描述和评价规则

多类标数据集的样本集合通常包括两类属性,一类是特征属性 $X = \{A_1, A_2, \cdots, A_r\}$,其中 $\{A_1, A_2, \cdots, A_r\}$ 分别代表样本的有限 r 个特征属性;另一类为类标属性 Y,该属性取值为多值,其值域离散有限,记为 $\{y_1, y_2, \cdots, y_q\}$. 多类标学习描述为: $T = \{(x_1, Y_1), (x_2, Y_2), \cdots, (x_m, Y_m)\}$. 其中: $x_j \in X$, $Y_j \subseteq Y$, $j = 1, 2, \cdots, m$; T 为训练数据集,生成一个多类标的分类器,利用该分类器对测试样本集 S 中各样本的类标集进行预测,得到预测类标集.

定义 6　$f(x_i, y_i)$ 为实例 x_i 可能属于类别 y_i 的概率.

定义 7　$\mathrm{rank}_f(x_i, y_i)$ 为实例 x_i 的类标集中 y_i 的排名指数.

对于一个多类标测试集 $D = \{(x_1, Y_1), (x_2, Y_2), \cdots, (x_m, Y_m)\}$,主要有以下评价标准[9].

定义 8　汉明距离差表示基于实例的类标被错误预测的次数占总预测次数的比重,即

$$\mathrm{hloss}(h) = \frac{1}{m}\sum_{i}^{m}\frac{1}{Q}|h(x_i)\Delta Y_i)|, \quad (1)$$

其中 Δ 表示类标被错误预测的基础.

定义 9　第一类标错误率表示分类器预测的每个实例优先级最高的类标不属于实例原有类标的次数占所有预测次数的百分比,即

$$\mathrm{one\text{-}error}_s(f) = \frac{1}{p}\sum_{i}^{p}\|[\arg\max_{y\in Y} f(x_i, y) \notin Y_i]\|. \quad (2)$$

定义 10　类标覆盖率表示属于测试集而没有被分类器正确预测的类标在所有预测类标中占的百分比,即

$$\mathrm{coverage}(f) = \frac{1}{p}\sum_{i}^{p}\|[\max_{y\in Y}\mathrm{rank}_f(x_i, y) - 1]\|. \quad (3)$$

定义 11　类标排名错误率根据分类器预测的类标优先级排名与实际类标优先级排名进行比较,看优先级排名发生了错误的类标在所有预测类标中占的百分比,即

$$\mathrm{rloss}_s(f) = \frac{1}{p}\sum_{i}^{p}\frac{1}{|Y_i||\overline{Y}_i|}|\{(y_1, y_2)|f(x_i, y_1) \le f(x_i, y_2), (y_1, y_2) \in Y_i \times \overline{Y}_i\}|. \quad (4)$$

定义 12　平均精度表示有分类器预测的类标,其优先级排名大于属于实例的原类标,且属于预测实例的原类标集的类标在所有预测类标中占的百分比,即

$$\mathrm{avgprec}_s(f) = \frac{1}{p}\sum_{i}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}(|\{y' | \mathrm{rank}_f(x_i, y') \le \mathrm{rank}_f(x_i, y), y' \in Y_i\}|/\mathrm{rank}_f(x_i, y)). \quad (5)$$

3 多类标问题分析

现有的关联分类算法一般包括挖掘规则阶段和建立分类器阶段. 在挖掘规则阶段中,先将训练集输入,通过算法挖掘出频繁项集以产生分类关联规则;然后通过一定的方法来筛选规则,保留那些分类效果最好的规则,输出到分类器中. 现有的关联分类算法按照以上步骤进行多类标学习主要存在以下两个难点:

1) 多类标规则. 传统的关联分类算法只能生成包含单个类标的规则,如果简单地将多类标考虑为单类标,结果将出现大量稀有类,而这些类的数据将很难有高于支持度阈值的规则对其分类. Kazawa[14] 使用这种将多类标考虑为一个新的单类标的方法,通过处理由海量的可能类导致的数据稀疏(Q 个主题将生成 2^Q 个类),将标记包含进一个相似性向量空间,将其中类似标记的原型向量放置得很靠近. 但是这种算法开销很大,运行效率太低.

2) 高维数据. 一般基于多类标的数据均属于高维数据,例如图片数据一般都能提取几百个有效特征. 如果使用一般的关联分类算法来挖掘规则,一是挖掘效率低,二是产生的规则数目惊人,可以产生 2^n ($n > 100$) 个规则数目. 而且因为数据所占空间随着维度增加而变得越来越稀疏,产生的关联规则将很难对数据进行正确的分类预测,从而出现维灾难[13].

针对以上难点,利用多类标 FP-tree 生成多类标规则;在筛选规则阶段,训练数据遍历规则集,筛选出和训练数据的类标相似度最高的规则;在组合基分类器阶段,提取数据集中部分特征属性建立 K 个子集,训练这 K 个子集产生的 K 个基分类器,组合基分类器投票来预测测试集. 通过以上方法,提出新的算法 MLAC,来解决基于多类标数据的关联规则分类问题.

4 MLAC 算法设计与实现

4.1 多类标 FP 树

当前常见的关联规则分类算法如 CBA 一般使用 APRIORI 来挖掘频繁模式,但存在效率低且容易产生大量冗余模式. 本文借鉴 FP-tree,提出了一种多类标规则 FP-tree 算法. 主要通过节点合并来有效生成多类标规则,减少树节点,从而有效降低规则筛选中遍历的时间消耗.

假设数据集 D 为 i 行 j 列,X_i 为第 i 行的数据,x_{ij} 为属于 X_i 的第 j 列属性. 多类标规则 FP-tree 算法描述如下:

Step1: 将 D 扫描一遍,找出大于给定支持阈值的属性-值对集合 $F = \{\langle X_i, x_{ij}\rangle\}$;

Step2: 按支持度计数对 $F = \{\langle X_i, x_{ij}\rangle\}$ 降序排列;

Step3: 为 D 中每个记录按 F 中的顺序将 t 在 F 中出现的属性-值对插入 FP-tree;

Step4: 插入的属性-值对将 t 的类标号同时插入,如果类标数 > 1,则将类标拆分分别插入;

Step5: 在 FP-tree 上自底向上递归地产生支持度大于 $\text{minsup}(R) = S_{\min}/|D|$ 的频繁模式;

Step6: 频繁模式中属性相同的分支对类标进行合并,删除合并后为空的节点,按 $\text{conf}(R)$ 对类标进行降序排列;

Step7: 找出所有大于 $\text{Minconf}(R) = S_{\min}/N(x)$ 的规则.

4.2 基于规则筛选的分类器设计

生成多类标规则 FP-tree 后,需要对规则集进行有效筛选生成分类器. 常见的关联分类算法如 CBA,CMAR 等都是对训练数据和规则的属性子集与类标进行完全匹配. 在多类标数据中,如果简单地将多类标考虑为单类标进行完全匹配,不但会导致数据稀疏降低分类器的准确率,还会因为很多数据无法找到完全匹配的规则而造成最终分类器规则集的缺失.

根据文献[7]中提出的公式,综合考虑集合同一性和一致性的类标集相似度的评定方法,计算多类标数据的类标集相似度,并根据其进行规则匹配和筛选.

基于多类标规则筛选的算法描述如下:

Step1: 规则比较,$R_1 > R_2$ (R_1 比 R_2 级别高)

if $\text{conf}(R_1) > \text{conf}(R_2)$;

if else $\text{conf}(R_1) = \text{conf}(R_2)$ and $\text{sup}(R_1) > \text{sup}(R_2)$;

else $\text{conf}(R_1) = \text{conf}(R_2)$ and $\text{sup}(R_1) = \text{sup}(R_2)$ 但 R_1 左边的属性值对比 R_2 少;

Step2: 置训练数据集中每个规则的覆盖计数为 0;

Step3: While 训练数据集不空 do {

for 以数据排序来让每个测试集遍历规则列表 do{ R 的覆盖计数置 0,

While 规则数目不为空 do{

for 以级别降序数据来考察的规则集中的每个规则 R do{

if 测试数据的类标与规则类标完全符合 then 数据被这个规则覆盖,规则覆盖数 + 1,

else if 没有找到与数据类标完全符合的规则,then 从规则中寻找 Simllarity 最高的规则,

else if 出现 $\text{Sim}(D_i, R_j) = \text{Sim}(D_i, R_k)$ then 选

择优先级高的规则,

else 没有规则能覆盖此数据,跳出到下一条数据}};

Step4: if 规则的覆盖计数为 0,then 从规则集中删除该规则;

Step5: 剩余规则组成最终的基分类器.

4.3　属性特征子集提取

在实际学习问题中,样本的特征集存在这样的分割,例如在彩色图像处理中,特征集通常包括与像素点位置相关的特征子集,或者称为坐标特征子集或与像素点的颜色特征相关的颜色特征子集. 对于这类问题,目前通过某种算法利用特征集的天然分割特性设计学习系统,以提高识别和分类性能.

本文使用的图片和基因数据集中:下载的原始图片数据集已经根据颜色空间中 RGB 值等特征被分割成 9 个独立的特征子集; Yeast 基因数据集则利用 K 均值区别分析法将其属性分成 K 个独立子集.

区别分析过程描述如下:

Step1: 找出预测变量的线性组合,使变量的组间变异相对于组内变异的比值为最大,而每一个线性组合与先前已经获得的线性组合均不相关.

Step2: 检验各组的重心是否有差异.

Step3: 找出哪些预测变量有最大的区别能力.

Step4: 根据新受试者的预测变量数值,将该受试者指派到某一群体.

4.4　基于关联分类的组合方法

为了保证组合分类器的分类效果好于单个分类器,需要遵循两个原则:一是组合分类器中的基分类器产生的误差是不相关的;二是基分类器的分类效果要比随机预测的效果好.

假设有 25 个基分类器,其平均准确率为 0.67,组合分类器最终的平均精度能达到 $\mathrm{avgprec}_s(f) = 0.94$. 在设计组合方法时,首先利用提取属性集特征子集,产生一组具有不同属性集的关联规则分类器,这样能保证基分类器的相互独立性;然后在筛选分类器规则阶段,由于利用了相似度的概念,保证了基分类器中有足够有效的规则来覆盖测试数据集,且无法找到有效覆盖规则的数据将不参与组合投票,避免出现零覆盖率的情况发生. 假设一个多类标子集 (C_2, C_3) 被分别判断为 (C_1, C_2), (C_2, C_3), (C_3) 空集,空集不参与投票,预测类标 (C_1, C_2, C_3) 的概率分别为:33 %,67 % 和 67 %,那么最终组合投票生成的预测类标集也将成功预测为 (C_2, C_3).

基于多重关联规则的组合方法的描述如下:

1) D 表示原始训练数据集, K 表示基分类器的个数, T 表示测试数据集.

2) for $i = 1$ to k do,

由 D 创建训练集 D_i,

由 D_i 构建基分类器 C_i,

end for;

3) for 每一个检验记录 X 属于 T do,

类标记预测为空则不参与投票,

$C^*(x) = \mathrm{vote}(C_1(x), C_2(x), \cdots, C_k(x))$;

根据 $C_1(x), C_2(x), \cdots, C_k(x)$,计算 $f(x_i, y_i)$;

if $f(x_i, y_i) \geq 0.5$ then 类标 y_i 加入记录 X 的类标集,

else,类标 y_i 不属于记录 X 的类标集,

end for.

4) 输出 $f(x_i, y_i)$ 和 $\mathrm{rank}_f(x_i, y_i)$.

最终得到测试数据集 T 的预测类标集 Pre-label 和 $f(x_i, y_i)$ 的集合 Outputs. 其中 Pre-label 和 Outputs 均为 $m * n$ 的二维矩阵, m 为测试数据集类别的个数, n 为测试数据集的实例个数.

5　实　　验

通过将 MLAC 与常见的 4 种多类标分类算法 ML-KNN, Adaboost.MH, MIMLSVM, Rank-Svm 的实验结果进行比较,利用 5 个指标来分析最后的结果,以考察这些算法的性能和稳定性,并通过比较算法的运行时间来判断其运行效率. 以上常见的 4 种多类标分类算法均在 Matlab R2006 下实现,而 MLAC 算法在 eclipse3.2 下通过 Java 和 Matlab R2006 混合编程实现. 所有实验使用配置为 T2600 双核的处理器和 2G 内存电脑,操作系统为 Windows Server 2000.

5.1　自然景观数据分类

多类标自然景观图像数据集来源于 corel image collection,约 22 % 的图片属于多类标数据. 实验将使用两种方法对这些图片提取特征向量,以便观察不同特征提取对算法性能的影响,判断算法性能稳定性的优劣:1) SBN 方法[15]. 将每张图片分为 9 个区间 $(K = 9)$,每个区间均有 15 个特征;2) 首先将这些彩色图片从 RGB 转换到 CIE Luv 空间[16],通过计算欧几里德距离提取特征值,然后使用 7 ×7 的矩阵将每个图片分成 49 个区间,最后将每个图片转换成 49 ×3 ×2 = 294 维的向量矩阵.

实验采用十字交叉验证法,最终结果是 10 次实验结果的均值和协方差.

1) 使用 SBN 方法提取特征的实验结果,见表 1. 从表 1 可以看出,MLAC 算法在 5 个指标的实验结果中有 4 个指标排名第一.

2) 使用转换到 CIELuv 空间的方法提取特征的

实验结果,见表 2. 从表 2 可以看出,MLAC 算法在 5 个指标的实验结果中有 3 个指标排名第一,在 4 个算法中实验结果表现最好.

表 1 5 个指标的均值和协方差

Evaluation Criterion	MLAC (K = 9)	ML-KNN	Adaboost.MH	MIML-Svm
Hamming Loss	0.170 ± 0.009	0.205 ± 0.009	0.231 ± 0.018	0.196 ± 0.103
One-error	0.304 ± 0.030	0.401 ± 0.034	0.451 ± 0.046	0.368 ± 0.032
Coverage	1.043 ± 0.076	1.135 ± 0.093	1.258 ± 0.137	1.115 ± 0.122
Ranking Loss	0.227 ± 0.028	0.218 ± 0.021	0.250 ± 0.031	0.211 ± 0.023
Average Precision	0.765 ± 0.017	0.741 ± 0.023	0.708 ± 0.030	0.756 ± 0.022

表 2 5 个指标的均值和协方差

Evaluation Criterion	MLAC (K = 9)	ML-KNN	Adaboost.MH	Rank-Svm
Hamming Loss	0.171 ± 0.015	0.169 ± 0.016	0.193 ± 0.014	0.253 ± 0.055
One-error	0.275 ± 0.033	0.300 ± 0.046	0.375 ± 0.049	0.491 ± 0.135
Coverage	0.904 ± 0.102	0.939 ± 0.100	1.102 ± 0.111	1.382 ± 0.381
Ranking Loss	0.211 ± 0.029	0.168 ± 0.024	N/A	0.278 ± 0.096
Average Precision	0.814 ± 0.025	0.803 ± 0.027	0.755 ± 0.027	0.682 ± 0.093

5.2 酵母菌(Yeast) 基因数据分类

在 Yeast 基因数据中,每一个基因描述成由一个级联微阵列表达与其他基因关系的数据. 为了使数据集更容易在实验中处理,Elisseeff 等[6] 在对数据进行预处理后只保留了已知结构和功能的类别,这些类别一共分为 4 个不同的层次. 最终用于实验的 Yeast 数据一共包含 2417 个基因,每个基因拥有 103 个特征向量. 被保留的已知结构和功能的类别有 14 个,每个基因平均属于 4.24 ±7.57 个不同的类别.

Yeast基因实验仍然采用十字交叉法求均值与协方差,实验结果见表 3.

表 3 5 个指标的均值和协方差

Evaluation Criterion	MLAC (K = 10)	ML-KNN	Adaboost.MH	Rank-Svm
Hamming Loss	0.176 ± 0.019	0.194 ± 0.010	0.207 ± 0.010	0.207 ± 0.013
One-error	0.218 ± 0.024	0.230 ± 0.030	0.244 ± 0.035	0.243 ± 0.039
Coverage	6.567 ± 0.222	6.275 ± 0.240	6.390 ± 0.203	7.090 ± 0.503
Ranking Loss	0.288 ± 0.019	0.167 ± 0.016	N/A	0.195 ± 0.021
Average Precision	0.756 ± 0.018	0.765 ± 0.021	0.744 ± 0.025	0.749 ± 0.026

在 Yeast 基因实验中,MLAC 的 5 个指标实验结果有 2 个排名第一,在 4 个算法中仅略次于有 3 个指标排名第一的 ML-KNN 算法.

综合以上针对自然景观和 Yeast 基因数据集的实验结果,MLAC 算法的表现优于其他 4 个算法.

5.3 算法效率

通过比较自然景观 SBN 方法实验中 4 个算法的运行时间(见表 4) 来分析这些多类标算法的效率. 4 个算法中,ML-KNN 耗时最短,只有不到 24 s,因为 ML-KNN 属于 lazy 学习法,没有针对问题建模,与其他 3 个算法相比虽然节省了建模时间,但可解释性较差. 而 MLAC 算法和 ML-KNN 算法相比,耗时仅多出 1 s 左右,但其分类性能优于 ML-KNN,基于规则分类的特点也造成了 MLAC 的可解释性好于 ML-KNN.

表 4 4 个算法运行时间 s

算法名称	MLAC	ML-KNN	Adaboost.MH	MIML-Svm
运行时间	25.1175	23.6179	> > 1000000	2192.9531

6 结 论

针对多类标学习的问题,分析了关联分类算法和组合的方法,提出了一种基于关联规则的多类标算法 MLAC,利用多类标 FP-tree 来分解组合生成多类标规则,并通过组合多重关联规则分类器进行分类预测. 实验结果表明,MLAC 算法在性能和稳定性上优于现有的其他多类标分类算法. 进一步提高关联规则组合分类器的性能可以考虑规则投票加权和各个类别之间的联系,这也是下一步继续研究和探索的方向.

参考文献(References)

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. Proc of the ACM SIGMOD Conf on Management of Data. Washington, 1993: 207-216.

[2] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]. Proc of the KDD. New York, 1998: 80-86.

[3] Bojarczuk C C, Lopes H S, Freitas A A. Genetic programming for knowledge discovery in chest-pain diagnosis[J]. IEEE Engineering in Medicine and Biology Magazine, 2000, 4(19): 38-44.

[4] Li W, Han J, Pei K. CMAR: Accurate and efficient classification based on multiple class-association rules [C]. Proc of the 2001 Int Conf on Data Mining. San Jose, 2001: 369-376.

(下转第 582 页)

[8] 姚升保, 岳超源, 张鹏, 等. 风险型多属性决策的一种求解方法[J]. 华中科技大学学报, 2005, 33(11): 83-85.
(Yao S B, Yue C Y, Zhang P, et al. Method for solving multi-attribute decision-making problems with risk[J]. J of Huazhong University of Science and Technology, 2005, 33(11): 83-85.)

[9] 姚升保, 岳超源. 基于组合赋权的风险型多属性决策方法[J]. 系统工程与电子技术, 2005, 27(12): 2047-2050.
(Yao S B, Yue C Y. Method for multiple attribute decision-making under risk based on synthetic weighting [J]. System Engineering and Electronics, 2005, 27(12): 2047-2050.)

[10] 宋业新, 张曙红, 陈绵云. 基于模糊模式识别的时序混合多指标决策[J]. 系统工程与电子技术, 2002, 24(4): 1-4.
(Song Y X, Zhang S H, Chen M Y. Hybrid multiple criteria decision-making with time series base on fuzzy pattern recognition [J]. System Engineering and Electronics, 2002, 24(4): 1-4.)

[11] 夏勇其, 吴祈宗. 一种混合型多属性决策问题的TOPSIS法[J]. 系统工程学报, 2004, 19(6): 630-634.
(Xia Y Q, Wu Q Z. A technique of order preference by similarity to ideal solution for hybrid multiple attribute decision making problems [J]. J of Systems Engineering, 2004, 19(6): 630-634.)

[12] 徐泽水, 达庆利. 区间数排序的可能度法及其应用[J]. 系统工程学报, 2003, 18(1): 67-70.
(Xu Z S, Da Q L. Possibility of method for ranking interval numbers and its application[J]. J of Systems Engineering, 2003, 18(1): 67-70.)

[13] 徐泽水. 模糊互补判断矩阵排序的一种算法[J]. 系统工程学报, 2001, 16(4): 311-314.
(Xu Z S. Algorithm for priority of fuzzy complementary judgement matrix [J]. J of Systems Engineering, 2001, 16(4): 311-314.)

[14] 樊治平, 姜艳萍. 模糊判断矩阵排序方法研究的综述[J]. 系统工程, 2001, 19(5): 12-18.
(Fan Z P, Jiang Y P. An overview on ranking methods of fuzzy judgement matrix [J]. System Engineering, 2001, 19(5): 12-18.)

(上接第 578 页)

[5] Yin X, Han J. CPAR: Classification based on predictive association rules[C]. Proc of the 3rd SIAM Int Conf on Data Mining. San Francisco, 2003.

[6] Comite F D, Gilleron R, Tommasi M. Learning multi-label altenating decision tree from texts and data[C]. Lecture Notes in Computer Science. Berlin: Springer, 2003: 35-49.

[7] 李宏, 陈松乔, 赵蕊, 等. 多值属性多类标数据决策树算法研究[J]. 模式识别与人工智能, 2007, 21(6): 815-820.
(Li H, Chen S Q, Zhao R, et al. Research on multi-valued and multi-labeled decision tree [J]. Pattern Recognition and Artificial Intelligence, 2007, 21(6): 815-820.)

[8] Elisseeff A, Weston J. A kernel method for multi-labelled classification [C]. Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002: 681-687.

[9] Schapire R E, Singer Y. Boostexter: A boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2): 135-168.

[10] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[11] Fadi A Thabtah. Peter cowling and yonghong peng [C]. Proc of the 4th IEEE Int Conf on Data Mining. Brighton, 2004: 217-224.

[12] Zhang M L, Zhou Z H. 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.

[13] Pang Ning, Tan Michael, Steinbach Vipin Kumar. Introduction to data mining[M]. Addison: Wes-ley, 2005.

[14] Kazawa H, Izumitani T, Taira H, et al. Maximal margin labeling for multi-topic text categorization[C]. Advances in Neural Information Processing Systems 17. Cambridge: MIT Press, 2005: 649-656.

[15] Maron O, Ratan A L. Multiple-instance learning for natural scene classification[C]. Proc of the 15th Int Conf on Machine Learning. Madison, 1998: 341-349.

[16] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification [J]. Pattern Recognition, 2004, 37(9): 1757-1771.