基于稳健模糊粗糙集模型的多标记文本分类

张 晶[1] 李德玉[1,2] 王素格[1,2] 李 华[1]
(山西大学计算机与信息技术学院 太原 030006)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]

摘 要 针对多标记数据的不确定性以及噪声数据的存在,提出了一种新的多标记稳健模糊粗糙分类模型。该模型是处理单标记分类问题的k-mean稳健统计量模糊粗糙分类模型的扩展应用。对于每个待分类数据,首先根据相似性计算方法,得到它们相对于各标记的隶属度;然后根据隶属度定义待分类数据与各标记的相关度;最后为每一组相关度赋予合适的阈值,得到相关的标记集合。在3个标准多标记数据集和1个真实多标记文本数据集上的实验结果表明,对于多标记文本分类问题,所提模型在6个常用的多标记评测指标上较常用的ML-*k*NN和rank-SVM多标记学习方法具有更高的准确率。

关键词 模糊粗糙集,k-mean稳健统计量,隶属度,多标记学习
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.058

Multi-label Text Classification Based on Robust Fuzzy Rough Set Model

ZHANG Jing[1] LI De-yu[1,2] WANG Su-ge[1,2] LI Hua[1]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]

Abstract Owing to the uncertainty of multi-label data and noise data, a novel multi-label robust fuzzy rough classification model was proposed, which is an extension of k-mean robust statistics fuzzy rough classification model that is used to solve the single label classification problem. First, for each unlabeled instance, the membership with respect to each label was obtained by similarity measures. Second, according to the membership, the degree of correlation was defined. Finally, an appropriate threshold was given to demarcate the correlated and uncorrelated labels. The experimental results on three benchmark multi-label datasets and one actual multi-label datasets indicate that the proposed model is superior to ML-*k*NN and rank-SVM across six popular multi-label evaluation metrics.

Keywords Fuzzy rough set, k-mean robust statistics, Membership, Multi-label learning

1 引言

数据分类问题一直是数据挖掘领域的一项重要研究内容。在传统的监督学习问题中,一个对象通常只标记为一个类别,我们称之为单标记学习问题。如果一个对象同时被标记为多个类别,对应的分类问题被称为多标记分类问题。例如,一篇关于雾霾的新闻文档可能涉及多个主题,如环境、经济、大气污染等;一张关于风景的照片可能涵盖多个场景,如夕阳、高山、树林等;一首歌曲往往也会包含有多种情感,如平静、愉快等;一个基因有可能包含多种功能,如转运类、蛋白质合成类等。这些例子从标记任务的角度,均需要多标记学习方法标记其多种标签。目前,多标记学习的研究已被应用于文档自动分类[1]、多媒体信息研究[2]、情感分类[3]以及生物信息学[4]等多个领域,成为机器学习领域中一个研究热点[5,6]。

对于多标记分类,如果限定每个对象只对应一个概念标记,那么传统的两类或多类学习问题均可以看作是多标记学习问题的特例。由于多标记数据具有语义不明确性和噪声,传统的监督学习方法难以直接解决此类数据的分类问题。粗糙集理论[7]是处理不确定性知识的一种有效的数学工具,而从粗糙集发展而来的模糊粗糙集模型[8]却对噪声数据非常敏感。因此,基于数值型数据的不确定性和噪声问题的处理角度,胡清华等人[9]提出了一种基于k-mean稳健统计量的模糊粗糙集模型。但该方法仅针对单标记问题,从而限制了模糊粗糙集的应用范围。鉴于多标记问题中每个实例具有多种类别特性,可以将其看作是多标记数据隶属于每个类标记的不确定程度,即类别不确定性。样本 x 隶属于某一类的下近似

到稿日期:2014-09-23 返修日期:2014-11-23 本文受国家自然科学基金项目(61175067,61272095),山西省科技攻关项目(20110321027-02),山西省回国留学人员科研项目(2013-014)资助。

张 晶(1990-),女,硕士生,主要研究方向为数据挖掘;李德玉(1965-),男,教授,博士生导师,主要研究方向为人工智能等,E-mail:lidy@sxu.edu.cn(通信作者);王素格(1964-),女,教授,博士生导师,主要研究方向为自然语言处理、智能检索等;李 华(1978-),女,博士生,主要研究方向为粒计算与数据挖掘。

隶属度越大，说明该样本对这一类的隶属程度越大，那么，将样本 x 划分到这一类的可能性也就越大。因此，稳健的模糊粗糙集模型既可以处理不确定性问题，又可以消除噪声数据带来的标记误差。

本文针对具有不确定性以及噪声的多标记数据，提出一种基于 k-mean 稳健统计量模糊粗糙集的多标记分类模型，记为 ML-kFRS（A multi-label classification model based on the k-mean fuzzy rough set）。该模型采用不同的相似性计算方法，根据每个待分类数据相对于各标记的隶属度，定义该数据与各标记之间的相关度，并赋予每个待分类数据一组相关度阈值，从而得到相应的标记集合，完成多标记分类任务。在真实文本数据集上进行的实验验证了本文提出的方法可以有效提高多标记文本分类的性能和效率。

2 多标记文本分类方法

如引言部分所述，多标记学习方法已经被用于解决很多真实世界中存在的问题。对于文本分类中遇到的歧义性问题，研究学者从不同的角度出发，提出了多种多标记学习方法。R. E. Schapire 和 Y. Singer[1] 提出的 BoosTexter 方法，是对 AdaBoost 提升方法进行集成，改变训练过程中的训练示例以及概念标记的权重，从而将一篇文档划分到多个概念标记中，该算法实现过程中建模消耗较大，有可能出现过拟合问题。A. K. McCallum[10] 提出了一种 EM 和 Bayes 相结合的方法，使用混合模型表示文档类别，再利用 EM 对混合权值和每一类的混合成分中字的分布进行学习。Ueda 和 Saito[11] 通过假设多标记文本有一个特征字的混合出现在每个类的单标记文档中，在 bag-of-words 表示下，提出了两种产生式概率模型。然而，文献[10,11]中的多标记学习方法都是基于文档中出现的字的频率来进行学习，因此仅适用于解决文档分类问题。除此之外，现有的许多多标记文本分类算法都是由传统的机器学习方法扩展而来的。A. Elisseeff 和 J. Weston[4] 采用“间隔最大化”策略，通过最小化 RankingLoss 损失函数，提出了一种基于支持向量机的方法，以解决多标记文档分类问题。该算法继承了 SVM 算法运算复杂度高的特点。张敏灵和周志华等人通过改造传统的 BP 神经网络，提出了一种多标记反向传播算法[12]，用于处理文本分类问题。该算法通过引入一种考虑多标记数据特征的新的误差函数，使待分类数据的相关标记排在非相关标记之前，其训练复杂度相对较高。在传统的 k 近邻学习算法的基础上，张敏灵和周志华等人又提出了 ML-kNN 多标记学习算法[13]，该算法根据待分类样本的 k 个近邻类别标记出现的先验和后验概率，基于最大化后验概率的原则确定文档的概念标记集合；但是，该算法忽略了标记之间的相关性，可能造成待分类数据无类别标识的情况。Comité 等人[14] 利用 AdaBoost. MH 算法[1] 对多标记交替决策树进行训练，通过改造交替决策树使其可以处理多标记数据。近年来，也有不少研究学者们通过改进上述方法，加入标记相关信息以及类别层次信息来提高多标记学习系统的性能。

3 基于 k-mean 稳健统计量模糊粗糙集的多标记分类模型

3.1 k-mean 模糊粗糙分类模型

通过模拟人类的推理方式，结合粗糙逼近和模糊粒化两种互补的不确定性推理方法，文献[8]提出了模糊粗糙集理论。其基本思想是当等价关系使模糊集合的论域变得粗糙时，定义此模糊集合的相应上近似和下近似。研究学者在模糊等价关系的基础上，通过定义不同的模糊逻辑算子，提出了多种模糊粗糙集模型[15]。而基于 k-mean 稳健统计量的模糊粗糙集模型[9]在计算模糊下近似隶属度时，不是用最近的异类样本来计算，而是采用与待分类样本最近的 k 个样本的均值。k-mean 模糊粗糙分类模型[9]就是基于模糊下近似提出的，它具有更好的稳健性。

定义 1[9] 对于决策表 $DT=\langle U,C,D\rangle$，R 是特征子集 $B\subseteq C$ 诱导出的模糊相似关系，且 $R(x,y)$ 随着距离 $\Delta(x,y)$ 的增加单调减小。令 A_i 为第 i 类样本的集合，对于 $\forall x\in A_i$，k-mean统计量模糊粗糙集模型的下近似和上近似算子分别定义为：

$$\underline{R}_{k\text{-}mean}A_i(x)=\min_{y\notin A_{ik\text{-}mean}}\{1-R(x,y)\} \tag{1}$$

$$\overline{R}_{k\text{-}mean}A_i(x)=\max_{y\in A_{ik\text{-}mean}}\{R(x,y)\} \tag{2}$$

其中，$R(x,y)$ 表示样本 x 与 y 之间的相似度，那么，$1-R(x,y)$ 可以看作是样本之间的伪距离。

对于 k-mean 模糊粗糙分类模型，给定一个决策表和一个待分类样本 x，首先需要计算样本 x 相对于每个类的 k-mean 模糊下近似隶属度，则 x 的类别判定方式为：

$$class(x)=\arg\max_{1\leqslant i\leqslant q}\{\underline{R}_{k\text{-}mean}class_i(x)\} \tag{3}$$

其中，$\underline{R}_{k\text{-}mean}class_i(x)$ 是 x 对于类 $class_i$ 的 k-mean 模糊下近似隶属度，q 为总的类别数。

该分类模型针对的只是单标记数据，考虑到多标记数据的特点，本文对 k-mean 模糊粗糙分类模型进行改进，使其能够更好地处理多标记文本分类问题。

3.2 ML-kFRS 模型的构造

首先给出本文使用的一些多标记分类学习的符号表示：对于一个给定的输入数据空间 $\mathcal{X}$，其标记空间用 $\mathcal{Y}=\{y_1,y_2,\cdots,y_q\}$ 表示，q 为标记个数。多标记训练数据集用 $G=\{(x_i,Y_i)\mid 1\leqslant i\leqslant n\}$ 表示，其中，$x_i\in\mathcal{X}$，$Y_i\subseteq\mathcal{Y}$，x_i 代表输入空间 $\mathcal{X}$ 中的一个训练数据，Y_i 是 x_i 的真实标记集合，n 表示训练数据集 G 中对象的个数。

ML-kFRS 模型的基本思想为：基于 k-mean 稳健统计量模糊粗糙集模型，计算每个待分类数据相对于标记集合 Y 中每个标记的 k-mean 模糊下近似隶属度，隶属度越大，表示该数据对某些类的隶属程度越大，属于这些类的可能性也就越大。将隶属度范围定义在 0～1 区间，即对所有隶属度完成到 0～1 区间的映射，结果可看作是待分类数据与各个标记的相关程度。然后，在对每个待分类数据选取合适阈值的条件下，将标记空间划分为“相关”标记和“无关”标记两类，从而得到相应的相关标记集合。

具体分为两步：

(1)计算待分类数据 x 与各个标记之间的相关度(Relevance)，用符号 H 表示。

假定 G' 表示训练数据集 G 中不具有标记 y_j 的数据集。x 相对于标记集合 Y 中的某个标记集合 y_j 的 k-mean 模糊下近似隶属度 $\underline{R}_{k\text{-}mean}y_j(x)(1\leqslant j\leqslant q)$ 取决于 x 与数据集 G' 之间的 k-mean 最近距离。

假设 x 与 G' 所有实例之间距离的顺序表示为 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_m\}$，$x$ 与 G' 之间的稳健 k-mean 最近距离表示为：

$$Dmin_{k\text{-}mean}(x,G')=\sum_{i=1}^{k}\theta_i/k \tag{4}$$

即为与 x 最近的 k 个异类样本的距离均值。

通常情况下，两个对象之间的相似性可以使用欧氏距离来度量。然而，在文本分类问题上，采用余弦相似度能更好地度量两文档之间的相似性[17]。因此，本文在文本数据集上采用余弦相似度计算隶属度，而对其他的数据集采用常用的欧氏距离计算隶属度。

当采用余弦相似度(Cosine Similarity)度量相似关系时，隶属度 $\underline{R}_{k\text{-}mean}y_j(x)_C$ 的计算可以表示为：

$$\underline{R}_{k\text{-}mean}y_j(x)_C=\min_{z\notin y_{jk\text{-}mean}}\{1-RC(x,z)\} \tag{5}$$

其中，$RC(x,z)$表示样本 x 和 z 之间的余弦相似关系。

当采用欧氏距离(Euclidean Distance)度量相似关系时，因为 $1-R(x,y)$可以看作是样本之间的伪距离，所以隶属度 $\underline{R}_{k\text{-}mean}y_j(x)_E$ 的计算可以表示为：

$$\underline{R}_{k\text{-}mean}y_j(x)_E=\min_{z\notin y_{jk\text{-}mean}}\{RE(x,z)\} \tag{6}$$

其中，$RE(x,z)$表示样本 x 和 z 之间的欧氏距离，可知 $\underline{R}_{k\text{-}mean}y_j(x)_E$ 等价于 x 与 G'之间的稳健 k-mean 最近距离。

由此，可以得到 x 相对于标记集合 Y 中每个标记的一组下近似隶属度值，全部用数组 $Degree_x$ 表示。

由于余弦相似度计算的一组隶属度的结果是在 0～1 范围内的，因此，相关度 H 可以直接用隶属度表示，即

$$H_C=Degree_x \tag{7}$$

由于欧氏距离计算的一组隶属度的结果不在 0～1 的范围内，因此需要将其映射到 0～1 区间，相关度 H 表示为：

$$H_E=\frac{Degree_x(k)-\min\{Degree_x\}}{\max\{Degree_x\}-\min\{Degree_x\}} \tag{8}$$

其中，k 表示数组中元素标号，$1\leqslant k\leqslant q$。

(2)选取合适的相关度阈值(Threshold)，用符号 δ 表示。

由于每个测试数据对应于各个标记的相关程度不同，阈值的选取需要具备合理性。如果给定一个固定阈值，阈值选取过大，可能导致应该被预测为相关标记的类却没有被分类器识别；反之，如果阈值选取过小，可能导致待分类数据被预测到了多个不相关标记。

例如，对于一个相关度集合 $H_i=\{0.89,0.85,0.5,0.7,0.6,0.8\}$的待分类数据 x_i，其真实标记集合 $Y_i=\{y_1,y_2,y_6\}$，当设定阈值 $\delta=0.9$ 时，分类器预测该数据的标记时，$h(x_i)=\emptyset$；当设定 $\delta=0.5$ 时，$h(x_i)=\{y_1,y_2,y_3,y_4,y_5,y_6\}$，显然与真实标记差距很大。

因此，考虑到每个数据与各个标记之间的相关度不同的情况，选取待分类数据 x_i 的所有标记相关度的数学期望作为划分标记的阈值，以增强算法的自适应能力。此时，$\overline{H}_i=0.72$，$h(x_i)=\{y_1,y_2,y_6\}$，可见与真实标记集合相符。用公式表示为：

$$\delta=\frac{1}{q}\sum_{i=1}^{q}H_i \tag{9}$$

那么，待分类数据 x 的标记集合的定义可以表示为：

$$Labelset(x)=\arg_{1\leqslant k\leqslant q}\{H(k)\mid H(k)\geqslant\delta\} \tag{10}$$

算法 1 给出了 ML-kFRS 分类算法的具体步骤。设 y_k 为标记空间 Y 中的某一个标记，其中 $1\leqslant k\leqslant q$，根据是否具有标记 y_k，可以将训练集 G 划分为具有标记 y_k 和不具有该标记两部分，前者记为 G_k，后者记为 G_k'。

算法 1 ML-kFRS 分类算法

输入：多标记训练数据集 G，近邻统计量参数 k，待分类数据 t，相应的标记集合 Y

输出：待分类数据 t 的预测标记集合 Y_t 以及与每个标记之间的相关度 H_t

步骤 1 初始化隶属度数组 $Degree_t=\emptyset$，相关度数组 $H_t=\emptyset$；

步骤 2 计算待分类数据 t 与训练集 G 中每个对象之间的距离(或者由相似度计算的伪距离)，得到距离数组 DIS_t；

步骤 3 for each $y_k\in y$

步骤 4 假设待分类数据 t 具有标记 y_k，t 与数据集 G_k'中每个对象之间的距离 DIS_{tk}'可从步骤 2 中提取；

步骤 5 对数组 DIS_{tk}'中的元素进行排序；

步骤 6 根据式(4)，结合有序的 DIS_{tk}'，计算 t 与数据集 G_k'之间的 k-mean最近距离；

步骤 7 根据式(5)或式(6)，计算待分类数据 t 相对于标记 y_k 的 k-mean 模糊下近似隶属度 $Degree_t$；

步骤 8 end for

步骤 9 根据式(7)或式(8)，计算待分类数据 t 与 q 个标记之间的相关度 H_t；

步骤 10 根据式(9)计算相关度阈值 δ_t；

步骤 11 根据式(10)得到待分类数据 t 的相关标记集合。

上述 k-mean MLFRS 算法中，步骤 2－步骤 8 计算待分类数据 t 相对于标记集合 Y 中每个标记的 k-mean 稳健模糊下近似隶属度。步骤 9 得到 t 相对于每个标记的相关度，可实现标记的排序。步骤 10、步骤 11 选取合适的阈值确定待分类数据 t 的标记集合。

当 $k=1$ 时，k-mean 最近距离等于最近邻两个样本间的距离，此时稳健模糊粗糙集模型退化为模糊粗糙集模型，退化后的 ML-kFRS 模型可以记为模糊粗糙集多标记分类模型(ML-FRS)，用于对比实验。

4 实验

4.1 数据集

本文一部分实验是在各个领域中常用的多标记数据集上的对比；另一部分针对多标记本文分类问题，在真实网页文档分类数据集上完成。两部分实验数据具体描述为：

(1)各个领域的多标记数据集

Emotion 数据集[3]：关于乐曲情感的多标记数据集。通过对某些音频剪辑、处理后得到。对象为 593 首乐曲，相关的 6 种情感为 6 个标记。

Image 数据集[13]：关于自然场景分类的多标记数据集。包含 2000 幅自然景色图片，每个对象被标记有 5 种图像类别(即相关标记)中的一个或多个。

Yeast 数据集[4]：关于预测酵母菌基因功能的多标记数据集。经过微阵列实验表示后，得到多种酵母菌细胞基因，包含了 2417 个基因对象，每个基因由 14 种功能类别(即相关标记)中的一种或几种组成。

(2)真实网页文档分类为标记数据集

Textdata 数据集：人工从网络上搜集的博客、新闻等文档构成的 1000 篇文本语料，其中包含 10 个类别，分别为：科技、体育、经济、军事、国际、政治、健康、饮食、电子、娱乐。每篇文档至少包含两个类别信息。经整理后，采用 DF 特征选择方法，选取词频较高的特征词构成特征词集；采用 TF-IDF 权重表示方法，将文本数据表示成向量空间模型的形式。

数据集的详细信息如表 1 所列。

表 1　4 个多标记数据集

Name	Instances	Labels	Features	Cardinality
Emotion	593	6	72	1.87
Image	2000	5	294	1.24
Yeast	2417	14	103	4.24
Textdata	1000	10	400	2.14

其中，Cardinality 指标记集合的势，表示平均每个数据具有的标记个数。

实验中，将 4 个原有数据集中的训练集和测试集混合，对其随机采样，采取 5 次交叉的方式完成实验。

4.2　评价指标

在多标记分类中，由于每个样本同时标有多个标记，因此其性能评价指标与传统的单标记分类相比较为复杂。本文采用下面的 6 种性能评价指标[18]从分类器预测性能和标记排序性能两方面反映出分类器分类效果和泛化性能。

假设一个多标记测试数据集 $S=\{(x_i,Y_i)\mid 1\leqslant i\leqslant p\}$，其中，$p$ 为测试样本数，$x_i\in\mathcal{X}$，$Y_i\subseteq\mathcal{Y}$。由分类器 h 预测到的对象 x_i 的标记集合记作 Z_i，表示为一个 q 维的二值向量。

在分类器预测性能方面，相关的两种评价指标[5]定义如下。

(1)Accuracy(AC)：通过对每个对象的真实标记集合 Y_i 和预测标记集合 Z_i 分别取交集与并集，以其大小的比值的宏平均来度量分类性能。其表示形式如下：

$$AC=\frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i\cap Z_i|}{|Y_i\cup Z_i|} \tag{11}$$

其取值越大，表示分类性能越优。

(2)F_1：准确率(Precision)和召回率(Recall)的一种权衡表示，常用于信息检索领域，经过改进之后也可以用于多标记的评估。其表示形式如下：

$$F_1=\frac{1}{p}\sum_{i=1}^{p}\frac{2|Y_i\cap Z_i|}{|Y_i+Z_i|} \tag{12}$$

该指标取值越大，表示分类性能越优。

在标记排序性能方面，各指标是基于每种算法的实值输出函数 $f:\mathcal{X}\times\mathcal{Y}\to\mathcal{R}$ 定义的。$f(\cdot,\cdot)$可以被转化为一个排序函数 $rank_f(\cdot,\cdot)$，该排序函数将对应于标记集合 $\mathcal{Y}$ 的任意标记 $y\in\mathcal{Y}$ 的输出 $f(x_i,Y_i)$映射到集合上$\{1,2,\cdots,q\}$，于是，如果 $f(x_i,y_1)>f(x_i,y_2)$，那么，$rank_f(x_i,y_1)<rank_f(x_i,y_2)$。$Y_i$ 表示对象 x_i 对应的标记集合。具体的评测指标[13]有：

(3)Average Precision(AP)：计算排序位于某个标记 $y\in\mathcal{Y}$ 之前的标记与对象 x_i 相关的比率。

$$AP(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|\{y'\mid rank_f(x_i,y')\leqslant rank_f(x_i,y),y'\in Y_i\}|}{rank_f(x_i,y)} \tag{13}$$

该值越大，表示分类器性能越优。

(4) Ranking Loss(RL)：没有被正确排序的标记对的平均比率。

$$RL(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{\{(y_1,y_2)\mid f(x_i,y_1)\leqslant f(x_i,y_2),(y_1,y_2)\in Y_i\times\overline{Y}_i\}}{|Y_i||\overline{Y}_i|} \tag{14}$$

其中，$\overline{Y}_i$ 表示 Y_i 在 Y 中的补集。该指标越小，表示分类器性能越优。

(5)One-error(OE)：计算在预测标记有序列中排名第一的标记与实际对象不相关的次数。

$$OE(f)=\frac{1}{p}\sum_{i=1}^{p}[\arg\max_{y\in Y}f(x_i,y)\notin Y_i] \tag{15}$$

其中，对于$[\pi]$，当 π 成立时，其值为 1，否则为 0。$OE(f)$的值越小，分类器的性能越优。

(6)Coverage(CV)：计算平均意义下找到一个对象的所有相关标记，需要遍历的预测标记有序列中的标记个数。

$$CV(f)=\frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i}rank_f(x_i,y)-1 \tag{16}$$

该指标的值越小，表示分类器性能越优。

4.3　实验结果及分析

4.3.1　统计量参数选择

k-mean 稳健统计量模糊粗糙集中近邻统计量 k 的选取决定了模型的抗噪能力。本次验以平均分类精度为标准选取近邻统计量 k 的值。图 1 给出了以数量 5 为间隔，参数 k 的取值在 1～40 时，3 个各领域常用数据集的平均分类精度 AP 的变化范围。其中，$k=1$ 时，稳健模糊的粗糙集模型退化为模糊粗糙集模型，此时不具备抗噪能力。

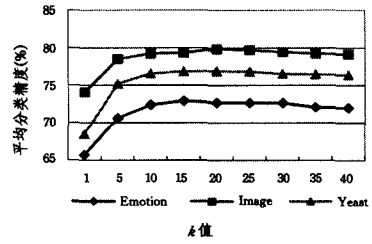

图 1　参数 k 取不同值时各数据集的平均分类精度

由图 1 可以看出，对于数据集 Emotion、Image、Yeast，在 $k=20$ 时，基本都取得了最好的平均分类精度。因此，包括文本数据集 Textdata 在内的对比实验全部在 k 取值为 20 的基础上完成。

4.3.2　对比实验

为了评测 ML-kFRS 分类算法的性能，本文将该分类器与 ML-FRS 和 ML-kNN[13]、rank-SVM[4] 进行对比。ML-kNN 和 rank-SVM 两种分类器在相关研究[12,16]中常被用于对比实验。ML-kNN 算法中，参数 k 设置为 10，平滑参数选用默认值；rank-SVM 算法中，采用线性核函数，其他均采用默认参数。符号↑表示指标的值越大分类性能越优；符号↓表示指标的值越小分类性能越优。分类结果汇总见表 2－表 5(加粗部分标注出各评价指标中的性能最优值)。

表 2　Emotion 数据集的实验结果(平均值±标准差)

Algorithm	AP↑	RL↓	OE↓	CV↓	AC↑	F_1↑
ML-kNN	0.7091±0.0210	0.2602±0.0270	0.3880±0.0388	2.2705±0.2099	0.3415±0.0428	0.4184±0.0431
Rank-SVM	0.6822±0.0170	0.2942±0.0155	0.4688±0.0418	2.3478±0.0994	0.3465±0.0231	0.4216±0.0272
ML-FRS	0.6570±0.0211	0.5309±0.0192	0.4082±0.0425	2.7372±0.1736	0.3927±0.0193	0.4689±0.0228
ML-kFRS	**0.7267±0.0284**	**0.2448±0.0211**	**0.3543±0.0603**	**2.1776±0.1558**	**0.4526±0.0221**	**0.5604±0.0213**

表 3 Image 数据集的实验结果(平均值±标准差)

Algorithm	AP↑	RL↓	OE↓	CV↓	AC↑	F_1↑
ML-kNN	0.7920±0.0158	0.1762±0.0196	0.3155±0.0195	0.9785±0.0864	0.5001±0.0252	0.5282±0.0276
Rank-SVM	0.7192±0.0677	0.2360±0.0729	0.4475±0.1111	1.1955±0.2789	0.4040±0.0834	0.4516±0.0905
ML-FRS	0.7405±0.0183	0.4090±0.0237	0.3465±0.0271	1.2790±0.0973	0.5596±0.0239	0.5946±0.0241
ML-kFRS	**0.7985±0.0154**	**0.1704±0.0178**	**0.3080±0.0213**	**0.9530±0.0825**	**0.5618±0.0145**	**0.6476±0.0145**

表 4 Yeast 数据集的实验结果(平均值±标准差)

Algorithm	AP↑	RL↓	OE↓	CV↓	AC↑	F_1↑
ML-kNN	0.7634±0.0123	**0.1684±0.0080**	0.2301±0.0161	6.2984±0.1369	0.5063±0.0152	0.6121±0.0138
Rank-SVM	0.7636±0.0155	0.1670±0.0096	0.2193±0.0235	6.3286±0.1506	0.5203±0.0156	0.6327±0.0132
ML-FRS	0.6847±0.0079	0.4300±0.0093	0.2462±0.0118	8.2698±0.1906	0.4760±0.0136	0.5713±0.0112
ML-kFRS	**0.7689±0.0125**	0.1687±0.0091	**0.2186±0.0205**	**6.2980±0.1728**	**0.5561±0.0148**	**0.6620±0.0127**

表 5 Textdata 数据集的实验结果(平均值±标准差)

Algorithm	AP↑	RL↓	OE↓	CV↓	AC↑	F_1↑
ML-kNN	0.9148±0.0150	0.0494±0.0084	0.0870±0.0222	1.7740±0.1076	0.7477±0.0127	0.7968±0.0131
Rank-SVM	0.9163±0.0140	0.0534±0.0101	0.0610±0.0192	1.8690±0.1240	0.6786±0.0228	0.7472±0.0179
ML-FRS	0.8967±0.0160	0.1448±0.0210	0.0760±0.0178	2.3440±0.1961	0.8183±0.0228	0.8542±0.0200
ML-kFRS	**0.9608±0.0125**	**0.0239±0.0093**	**0.0310±0.0147**	**1.4120±0.1080**	**0.8543±0.0158**	**0.8977±0.0149**

由表 2—表 4 给出的 4 种算法在 3 个数据集上的对比结果可以看出,本文提出的 ML-*k*FRS 算法在绝大多数指标上都表现出明显的优势;尤其在分类器预测性能方面,Emotion、Image、Yeast 3 个数据集在 F_1 上,相较于 ML-*k*NN 和 rank-SVM 两者结果的最优值,分别提高了 13.88%、11.94%、4.99%,说明 ML-*k*FRS 算法在预测标记方面有很高的准确率,这有利于只需获得标记结果的信息检索过程。表 5 给出的结果显示,对于多标记文本数据集 Textdata,采用余弦相似性的 ML-*k*FRS 分类算法在各个指标上的优势较其他 3 个数据集更加明显;与 ML-FRS 算法分类结果相比,ML-*k*FRS 算法在各个指标上均有很大的提高,说明其对于多标记数据具有很好的抗噪效果。此外,ML-*k*NN 算法对于每条数据的标记预测是独立完成的,少部分测试数据最终的分类结果可能会出现没有标识的情况,这对于多标记分类是无意义的。而由第 3 节构建的 ML-*k*FRS 分类模型可以看出,本文算法不会出现这种情况。

显著性水平为 5%的配对 t 检验结果表明,对于数据集 Emotion,ML-*k*FRS 算法在指标 Accuracy 和 F_1 上的性能显著优于 ML-*k*NN 算法,在所有的 6 个指标上都显著优于 rank-SVM 和 ML-FRS。对于数据集 Image,ML-*k*FRS 算法在指标 Accuracy 和 F_1 上的性能显著优于 ML-*k*NN 算法,在指标 Average Precision、Accuracy 和 F_1 上显著优于 rank-SVM 算法,在除 Accuracy 外的其他 5 个指标上显著优于 ML-FRS。对于数据集 Yeast,ML-*k*FRS 算法在指标 Average Precision、Accuracy 和 F_1 上的性能显著优于 ML-*k*NN 算法,在指标 Accuracy、F_1 上显著优于 rank-SVM 算法,在所有的 6 个指标上均优于 ML-FRS。对于数据集 Textdata,ML-*k*FRS 算法在所有 6 个指标上均显著优于其他 3 种方法,证明了该算法对于多标记文本分类的有效性。

4.3.3 时间分析

为了更加具体地展现 ML-*k*FRS 分类模型的优势,除了考虑实验结果的好坏外,时间与空间复杂度也应该考虑。本实验是对 ML-*k*FRS 算法以及 ML-*k*NN 和 Rank-SVM 3 种算法的分类时间进行分析。表 6 给出了 Emotion、Image、Yeast、Textdata 4 个数据集上各算法的时间消耗。

表 6 3 种算法在所有数据集上的平均分类时间(s)

Datasets	ML-kNN	Rank-SVM	ML-kFRS
Emotion	0.66	134.13	0.30
Image	12.80	170.58	4.54
Yeast	12.12	4693.52	4.01
Textdata	3.85	466.94	2.15

由表 6 可以看出,ML-*k*FRS 算法分类的时间消耗是最小的,ML-*k*NN 算法分类时间居中,而 Rank-SVM 算法是最费时的。与大部分分类算法相似,ML-*k*NN 和 Rank-SVM 算法都需要消耗一定的时间用于训练,尤其是 rank-SVM 算法,较不需要训练过程的 ML-*k*FRS 算法,其复杂度明显偏高。因此,本文提出的 ML-*k*FRS 分类算法在分类时间上表现出很大的优势。

结束语 针对多标记数据中不确定性以及噪声数据的存在,本文基于 k-mean 稳健统计量模糊粗糙集模型,构建了一种基于 k-mean 稳健统计量模糊粗糙集的多标记分类模型(ML-*k*FRS),其能够有效解决多标记分类问题。在多个真实多标记数据集上进行的实验,不仅验证了该算法对于多标记学习(尤其是文本分类)的有效性,而且证明了该算法在分类时间上具有较高的效率。

本文中,每个测试数据标记集合的预测以及与各标记之间相关性的确定,只是由该测试数据相对于不同类的隶属程度决定,忽略了多标记数据中标记之间的相关性。因此,通过加入标记间的相关性来提高分类器的准确性,是今后需要研究的工作。

参 考 文 献

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

[2] 郝虹,计华,张化祥,等. 基于 I2C 距离和标记相关性的多标记场景分类[J]. 计算机科学,2014,41(1):88-90

Hao Hong, Ji Hua, Zhang Hua-xiang, et al. Multi-label scene

classification based on I2C distance and label dependency[J]. Computer Science,2014,41(1):88-90

[3] Trohidis K,Tsoumakas G,Kalliris G,et al. Multi-label classification of music into emotions[C]// Proceeding of 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia,PA,USA,2008:325-330

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

[5] Tsoumakas G,Katakis I. Multi-Label Classification: An Overview[J]. International Journal of Data Warehousing and Mining,2007,3(3):1-13

[6] Zhang Min-ling, Zhou Zhi-hua. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(8):1819-1837

[7] Pawlak Z. Rough Sets [J]. International Journal of Computer and Information Sciences,1982,11(5):341-356

[8] Dubois D,Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems,1990,17:191-208

[9] Hu Qing-hua, Zhang Lei, An Shuang, et al. On robust fuzzy rough set models[J]. IEEE Transactions on Fuzzy systems, 2012,20(4):636-651

[10] McCallum A. Multi-label text classification with a mixture model trained by EM[C]// Proc of Working Notes of the AAAI'99 Workshop on Text Learning. Menlo Park, CA: AAAI Press, 1999

[11] Ueda N,Saito K. Parametric mixture models for multi-label text [C]// Advances in Neural Information Processing Systems 15. Cambridge,MA:MIT Press,2003:721-728

[12] Zhang Min-ling, Zhou Zhi-hua. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Transactions on Knowledge and Data Engineering, 2006,18(10):1338-1351

[13] Zhang Min-ling, Zhou Zhi-hua. ML-*k*NN: A lazy learning approach to muti-label learning[J]. Pattern Recognition,2007,40 (7):2038-2048

[14] Comité F D,Gilleron R,Tommasi M. Learning multi-label alternating decision tree from texts and data[C]// Lecture Notes in Computer Science 2734. Berlin:Springer,2003:35-49

[15] Yeung D S,Chen D G,Tsang E C C,et al. On the Generalization of Fuzzy Rough Sets[J]. IEEE Transactions on Fuzzy Systems, 2005,13(3):343-361

[16] 郑伟,王朝坤,刘璋,等. 一种基于随机游走模型的多标记分类算法[J]. 计算机学报,2010,33(8):1418-1426
Zheng Wei,Wang Chao-kun,Liu Zhang,et al. A multi-label classification algorithm based on random walk model[J]. Chinese Journal of Computers,2010,33(8):1418-1426

[17] 广凯,潘金贵. 一种基于向量夹角的 k 近邻多标记文本分类算法[J]. 计算机科学,2008,35(4):205-207
Guang Kai,Pan Jin-gui. An *k*NN algorithm based on vector angle for multi-label text categorization[J]. Computer Science, 2008,35(4):205-207

[18] Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data [M]// Data Ming and Knowledge Discovery Handbook. Berlin: Springer,2010:667-685

(上接第 261 页)

[9] Yu Hong-liang, Zheng Dong-dong, Zhao B Y, et al. Understanding User Behavior in Large-scale Video Streaming Services [J]. Computer Science Department,2006,40(4):18-21

[10] Riad A M, Elmogy M, Schehab A I. A Framework for Cloud P2P VoD System based on User's Behavior Analysis[J]. International Journal of Computer Applications,2013,76(6):20-26

[11] He Yi-feng, Shen Guo-bin, Xiong Yong-qiang. Optimal Prefetching Scheme in P2P VoD Applications With Guided Seeks[J]. IEEE Transaction on Multimedia,2009,11(2):138-151

[12] 王庆波,代亚非,田敬,等. 基于特征信息定位的 P2P 网络模型: Barnet[J]. 软件学报,2003,14(8):1481-1488
Wang Qing-bo,Dai Ya-fei,Tian Jing,et al. An Infrastructure for Attribute Addressable P2P Network: Barnet[J]. Jornal of SoftWare,2003,14(8):1481-1488

[13] 张玉洁,何明,孟祥武. 基于用户需求的内容分发点对点网络系统研究[J]. 软件学报,2014,25(1):98-117
Zhang Yu-jie,He Ming,Meng Xiang-wu. Research on CDN-P2P system over user requirements[J]. Journal of Software,2014,25 (1):98-117

[14] Sutton R S,Barto A G. Reinforcement Learning—An Introduction[M]. MIT Press,1998

[15] Tanner B,Sutton R S. Temporal-Difference Networks with Eligibility Traces[C]// Proceedings of the 22nd International Conference on Machine Learning. 2005:1313-1320

[16] Murphy S A. A Generalization Error for Q-Learning[J]. Journal of Machine Learning Research,2005,6:1073-1097

[17] Grondman I,Busoniu L,Lopes G A D,et al. A Survey of Actor-Critic Reinforcement Learning: Stan-dard and Natural Policy Gradients[J]. IEEE Transactions on Systems, 2012, 42(6): 1291-1307

[18] Mahadevan S,Maggioni M. Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes[J]. Journal of Machine Learning Research, 2007,8(10):2169-2231

[19] Kwong K W,Tsang D H K. A Congestion_Aware Search Protocol for Unstructured Peer-to-Peer Networks[C]// Parallel and Distributed Processing and Applications: Second International Symposium(ISPA). HongKong,China,2005:319-329

[20] Huang Cheng,Li Jin,Ross K W. Can Internet Video-on-Demand be Profitable? [J]. ACM SIGCOMM Computer Communication Review,2007,37(4):133-144

[21] Abbasi U, Ahmed T. COOCHING: Cooperative Prefetching Strategy for P2P Video-on-Demand System[C]// Wired-Wireless Multimedia Networks and Services Management: IEEE International Conference on Management of Multimedia and Mobile Networks and Service(MIMNS 2009). 2009:195-200