第52卷第5期 南京航空航天大学学报 Vol. 52 No. 5
2020年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2020

DOI:10.16356/j.1005-2615.2020.05.013

基于聚类的多标记选择性集成

张佳欢[1],李磊军[2,3],李美争[1],米据生[2],解 滨[1]

(1.河北师范大学计算机与网络空间安全学院,石家庄,050024;2.河北师范大学数学科学学院,石家庄,050024;
3.河北师范大学数学博士后科研流动站,石家庄,050024)

摘要:多标记学习和选择性集成是机器学习中的两个热点研究问题。本文利用聚类思想探究多标记学习中的选择性集成,提出了两种具体的多标记选择性集成算法:基于最小距离的簇中心选择算法(Minimum distance based cluster center selection,MDCCS)和基于K-means的簇中心选择算法(K-means based cluster center selection,KMCCS)。在所提出的算法中,如何度量学习器之间的距离是其能否成功的关键因素。本文首先基于学习器的分类结果对其进行重新表示,在此基础上给出了学习器之间距离的计算方式。此外,对于算法中的空簇问题给出了两种解决方法。基于Mulan数据库中的多标记数据集和5种评价指标对所提算法进行了详细的分析,实验结果表明了所提算法的有效性。

关键词:选择性集成;多标记学习;聚类;机器学习

中图分类号:O236 **文献标志码:**A **文章编号:**1005-2615(2020)05-0768-09

Multi-label Selective Ensemble Based on Clustering

ZHANG Jiahuan[1], *LI Leijun*[2,3], *LI Meizheng*[1], *MI Jusheng*[2], *XIE Bin*[1]

(1. College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang, 050024, China;
2. College of Mathematics Science, Hebei Normal University, Shijiazhuang, 050024, China;
3. Postdoctoral Research Workstation of Mathematics, Hebei Normal University, Shijiazhuang, 050024, China)

Abstract: Multi-label learning and selective ensemble are two hotspot problems in machine learning. Selective ensemble is explored in multi-label learning based on clustering. Two multi-label selective ensemble algorithms, including minimum distance based cluster center selection (MDCCS) and K-means based cluster center selection (KMCCS), are proposed. The key is to measure the distance between base learners in the proposed algorithms. The learners are represented based on their classification results, then the distance between the learners can be calculated. Besides, two solutions are proposed to solve the problem of empty cluster in the algorithm. Based on the multi-label data sets in Mulan database and five evaluation indexes, the proposed algorithms are analyzed in detail. The experimental results show the effectiveness of the proposed algorithms.

Key words: selective ensemble; multi-label learning; clustering; machine learning

在现实生活中,样本通常具有多个语义标记[1-3],多标记学习近些年已经引起了广泛关注。迄今为止,学者们提出了一系列处理多标记数据的方法。例如,将多标记数据转化为单标记数据,然

基金项目:国家自然科学基金(61502144, 62076088, 61672206)资助项目;河北省自然科学基金(F2018205196, F2019205295)资助项目;河北省高等学校自然科学基金(BJ2019014)资助项目;河北省博士后择优资助科研基金(B2016003013)资助项目;河北省三三三人才工程培养经费(A2017002112)资助项目。

收稿日期:2020-06-06;**修订日期:**2020-07-11

通信作者:李磊军,男,副教授,E-mail:lileijun1985@163.com。

引用格式:张佳欢,李磊军,李美争,等.基于聚类的多标记选择性集成[J].南京航空航天大学学报,2020,52(5):768-776. ZHANG Jiahuan, LI Leijun, LI Meizheng, et al. Multi-label selective ensemble based on clustering[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5):768-776.

后利用已有的处理单标记数据的方法对多标记数据进行处理[1,4]；或者对已有的处理单标记数据的方法进行调整，使之适应多标记数据[5-6]。

另外，集成学习在机器学习领域同样已经引起了广泛的关注[7-9]，已经被广泛应用于解决各种问题[10-14]。集成学习虽可提供更强的泛化性能，但随着基学习器的增加，运算速度逐渐下降，同时所需存储空间逐渐增大。因此，有选择地集成部分基学习器能够加快预测速度并减少存储空间，更重要的是，理论分析和实验结果表明，它能够进一步提高原有集成系统的泛化能力，得到更好的预测效果。因此选择性集成[15]受到研究者们的关注，成为该领域的一个研究重点。

目前，选择性集成已经得到了广泛的研究。Chen 等提出了 D3C 算法，该算法构造了 K 均值聚类与动态选择循环框架，结合序列搜索方法选择学习器，多标记问题下的 D3C 算法可拆分为多个单标记问题分别求解[16]。Croux 等则利用包外错误率对学习器进行排序，截取包外错误率较小的部分基学习器作为选择结果[17]。Zhou 等提出了 GASEN 算法，其利用遗传算法对基神经网络的权重进行优化，依据相应权重对基神经网络进行选择集成[18]。Xing 等则提出了一种基于二阶瑞利熵和 $L1$ 范数的多样性测度的选择性集成方法[19]。Zhang 等基于概率粗糙集模型给出了一种多标记下的特征选择方法，并通过集成使得泛化性能进一步提升[20]。Wu 等构造了一组层次树，以层次的方式通过给出的标记依赖性识别相关标记，并对多个标记依赖性进行集成得到最终标记[21]。

本文首先介绍了 K-means 聚类算法等基础知识，进而提出了两种基于聚类的选择性集成算法。基于这两种算法进行了一系列实验，详细地分析了两种算法中的参数对多标记分类性能的影响。另外，对于算法中的空簇问题给出了两种解决方法。

1　多标记学习及 K-means 算法[22]

1.1　多标记学习

假设 $\boldsymbol{X}=\mathrm{R}^N$ 表示 N 维实数样本空间，其中 $U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，$L=\{\eta_1,\eta_2,\cdots,\eta_q\}$ 表示标记集合，$r=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)|i=1,\cdots,n\}$ 表示样本在标记上的映射关系。样本 $\boldsymbol{x}_i$ 为 N 维向量，$\boldsymbol{Y}_i=(y_{i1},y_{i2},\cdots,y_{iq})$ 表示样本 $\boldsymbol{x}_i$ 对应的标记集合，若样本 $\boldsymbol{x}_i$ 具有第 q 个标记，则 $y_{iq}=1$，否则 $y_{iq}=0$[5]。

给定测试集合 $T=\{(\boldsymbol{x}_i^{\sim},\boldsymbol{Y}_i^{\sim})|1\leqslant i\leqslant m\}$，$\boldsymbol{Y}_i^{\sim}\subseteq L$ 表示 $\boldsymbol{x}_i^{\sim}$ 的真实标记子集，$\boldsymbol{Y}_i'\subseteq L$ 是由多标记学习器预测的标记集。由标记排序方法对标记 η 的预测等级表示为 $\mathrm{rank}(\eta)$，最相关的标记获得最高等级 1，最不相关的最低等级为 q。为了评价学习器的预测性能，提出了以下 5 种评价指标[5]。

(1)平均精度(Average precision，AP)表示预测标记集合中的标记排序等级与实际标记集中的某个 $\eta\in\boldsymbol{Y}_i^{\sim}$ 的特定标记相同或者更高的等级对应的概率，实际反映了预测标记的平均准确率，其定义如下

$$\mathrm{AP}=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|\boldsymbol{Y}_i^{\sim}\right|}\times\sum_{\eta\in\boldsymbol{Y}_i^{\sim}}\frac{\left|\left\{\eta'\in\boldsymbol{Y}_i^{\sim}:\mathrm{rank}_i(\boldsymbol{x}_i^{\sim},\eta')\leqslant\mathrm{rank}_i(\boldsymbol{x}_i^{\sim},\eta)\right\}\right|}{\mathrm{rank}_i(\boldsymbol{x}_i^{\sim},\eta)}$$

(2)覆盖率(Coverage，CV)表示覆盖预测样本标记的平均距离，其定义如下

$$\mathrm{CV}=\frac{1}{m}\sum_{i=1}^{m}\max_{\eta\in\boldsymbol{Y}_i^{\sim}}\mathrm{rank}_i(\eta)-1$$

(3)排序损失(Ranking loss，RL)表示不相关标记比相关标记排序更高的次数，其定义如下

$$\mathrm{RL}=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|\boldsymbol{Y}_i^{\sim}\right|\left|\overline{\boldsymbol{Y}_i^{\sim}}\right|}\left|\{(\eta_1,\eta_2)\times\mathrm{rank}_i(\eta_1)>\mathrm{rank}_i(\eta_2),(\eta_1,\eta_2)\in\boldsymbol{Y}_i^{\sim}\times\overline{\boldsymbol{Y}_i^{\sim}}\}\right|$$

(4)最高标记错误率(One error，OE)表示预测的最高等级标记不在样本真实标记集合的次数，其定义如下

$$\mathrm{OE}=\frac{1}{m}\sum_{i=1}^{m}\delta\left(\arg\min_{\eta\in L}\mathrm{rank}_i(\eta)\right)$$

$$\delta(\eta)=\begin{cases}1 & \eta\notin\boldsymbol{Y}_i^{\sim}\\ 0 & \text{其他}\end{cases}$$

(5)汉明损失(Hamming loss，HL)表示由每个样本真实标记集与预测标记集不同的元素个数反映的损失，其定义如下

$$\mathrm{HL}=\frac{1}{m}\sum_{i=1}^{m}\frac{\left|\boldsymbol{Y}_i^{\sim}\oplus\boldsymbol{Y}_i'\right|}{\boldsymbol{q}}$$

式中：$\oplus$ 代表集合之间的对称差，即布尔运算中的逻辑异或运算。

1.2　K-means 算法

K-means 算法框架如下：给定训练样本集 $U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，划分所得簇为 $S=\{s_1,s_2,\cdots,s_c\}$，K-means 算法旨在最小化平方误差

$$\mathrm{loss}=\sum_{i=1}^{c}\sum_{x\in s_i}\left\|\boldsymbol{x}-\boldsymbol{\mu}_i\right\|_2^2$$

式中：$\boldsymbol{\mu}_i=\frac{1}{\left|s_i\right|}\sum_{j=1}^{|s_i|}\boldsymbol{x}_j$ 为某簇 s_i 的均值向量。该式可直观地反映出某簇内的所有样本距均值向量 $\boldsymbol{\mu}_i$

的远近程度，loss值越小则簇内样本在该标准下的相似度越高。

衡量样本间相似度通常采用的方法为计算样本间的距离。若样本间距离增大，则样本间相似度减弱。以下将介绍各种样本间距离的计算方法。

(1)p-范数

给定N维实数空间$\mathbf{R}^N$，并且假设$U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$为$N$维样本集合，其中$\Delta:\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{R}$属于$\mathbf{R}^N$上的一个度量值，度量空间上的距离可表为：$\Delta(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{k=1}^{N}\left|\boldsymbol{x}_{ik}-\boldsymbol{x}_{jk}\right|^p\right)^{1/p}$。$p=1$时，$\Delta$表示哈曼顿距离；$p=2$时，$\Delta$表示欧式距离；$p=\infty$时，$\Delta$表示切比雪夫距离。

(2)汉明距离

给定两个N维向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_N)^{\mathrm{T}}$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，则两个$N$维向量间的汉明距离定义为$\mathrm{HD}=\dfrac{\sum_{i=1}^{N}f(x_i\oplus y_i)}{N}$，其中$f(\cdot)=\begin{cases}1 & x_i\oplus y_i\text{为真}\\ 0 & x_i\oplus y_i\text{为假}\end{cases}$。

(3)余弦距离

给定两个N维向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_N)^{\mathrm{T}}$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，则两个$N$维向量之间的余弦距离定义为$\mathrm{COSD}=\dfrac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}$。

(4)杰卡德距离

给定两个N维向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_N)^{\mathrm{T}}$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，则两个$N$维向量间的杰卡德距离定义为$\mathrm{JD}=\dfrac{\sum_{i=1}^{N}(x_i\oplus y_i)}{N-\mathrm{allzero}}$，其中allzero为$x_i$与$y_i$同时为0的次数。

2 基于聚类的多标记选择性集成算法

假设H为基学习器集合：$H=\{h_1,h_2,\cdots,h_T\}$；U为N维样本集合：$U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{iN})$；$Y$为$q$维标记集合：$Y=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m\}$，其中$\boldsymbol{Y}_i=(y_{i1},y_{i2},\cdots,y_{iq})$，$y_{iq}=1$则表示样本$x_i$具有第$q$个标记，$y_{iq}=0$则表示样本$x_i$不具有第$q$个标记。

假设基学习器集合是已知的，任一基学习器均由全体样本的q个预测标记组成，其可表示为$n\times q$的0-1值矩阵。

给定任意两个基学习器h_i与h_j，则h_i与h_j之间的距离可表示为$D(h_i,h_j)=\dfrac{\sum_{k=1}^{n}f(h_{ik},h_{jk})}{n}$，其中$f(\cdot)$可通过1.2节中所给出的距离方式进行计算，$\boldsymbol{h}_{ik}$为第$i$个基学习器中第$k$个样本对应的$q$维预测标记向量，$n$为样本个数。$f(h_{ik},h_{jk})$表示基学习器$h_i$与$h_j$在第$k$个样本下的$q$维预测标记向量间的距离，$h_i$与$h_j$间的距离为所有样本下预测标记向量间距离累加和的平均值，该值越小则基学习器间的相似度越强。

以下将给出两种基于聚类的多标记选择性集成算法。

给定$H=\{h_1,h_2,\cdots,h_T\}$为基学习器集合，则基学习器集合H可通过寻找K个簇中心点$\{S_p\}_{p=1}^{K}$使得$\sum_{i=1}^{T}\sum_{j=1}^{K}\min D(h_i,h_j)$达到最小值，与此同时基学习器集合$H$被划分为$K$个簇。

为了寻找最佳簇数，K的初始值可设为1，并以步长为1增长，直至$\sum_{i=1}^{T}\sum_{j=1}^{K}\min D(h_i,h_j)$达到最小值。分析可知，若使该式达到最小值，则时间复杂度为$O(2^n)$。为了降低时间复杂度，本文给出了该式的近似求解方法。具体过程详见算法1。其中，last_dist表示上次的总距离，this_dist表示本次的总距离。numofcluster为簇的个数，步骤2假设以每一个基学习器距其他学习器的总距离越小为标准选择簇中心。

算法1 基于最小距离的簇中心选择算法(MDCCS)

输入：$H=\{h_1,h_2,\cdots,h_T\}$，last_dist$=\infty$，this_dist，numofcluster$=1$。

输出：$H_K=\{H_1',H_2',\cdots,H_K'\}$。

步骤1 $\forall h_i,h_j\in H$，计算$D(h_i,h_j)$

步骤2 for $i=1:T$

计算$\sum_{j=1}^{n}D(h_i,h_j)$;

end for

步骤3 对$D(h_i,\cdot)$由小到大进行排序

步骤4 while(簇数$\leqslant T$)

如果本轮簇划分的下的总距离this_dist小于上轮簇划分的总距离last_dist，则最佳簇数则可继续增加直至当前总距离最小，由此确定最佳簇数numofcluster;

end while

步骤5 返回$H_K=\{H_1',H_2',\cdots,H_K'\}$。

对于this_dist，首先确定本轮簇数，进而根据

确定的簇数选择簇中心，并依次计算每簇的总距离，最终各个簇的总距离累加可得 this_dist。

给定 $H=\{h_1,h_2,\cdots,h_T\}$ 为基学习器集合，则可基于 K-means 对基学习器集合进行聚类。初始簇中心可随机从 H 中任取 K 个，由于任意基学习器皆对应 0-1 值矩阵，则每轮迭代的质心可由某簇内的全部基学习器对应的 0-1 值矩阵求得，求得的质心可依据所得矩阵计算与任意学习器间的距离，直至上次计算结果与本次计算结果相同时停止迭代。具体过程详见算法 2。

算法 2　基于 K-means 的簇中心选择算法 (KMCCS)

输入：$H=\{h_1,h_2,\cdots,h_T\},K$

输出：$H_K=\{H_1',H_2',\cdots,H_K'\}$

步骤 1　从 H 中随机选择 K 个样本作为初始状态下的簇中心点，并记为 $h_1',h_2',\cdots,h_K'$。

步骤 2　计算各个簇中心 h_j' 与任意基学习器 h_i 之间的距离，并将 h_i 划分给距其最近的簇 H_j'。

步骤 3　簇划分确定后即可计算新的簇中心，由于任意基学习器对应 0-1 值的标记矩阵，则可对任意簇内全部基学习器的 0-1 值矩阵的和取平均值作为新的簇中心。

步骤 4　重复步骤 2、3，直至本次簇划分与上次簇划分相同则可停止迭代。

步骤 5　返回 $H_K=\{H_1',H_2',\cdots,H_K'\}$。

3　实验分析

本文的所有多标记数据集来源于 Mulan 数据库，具体描述如表 1 所示。其中，Emotions[23] 数据集是音乐领域的数据集，总计 593 个实例，其中特征个数为 72，标记个数为 6；Birds[24] 为音频领域的数据集，总计 645 个实例，其中特征个数为 260，标记个数为 19；Flags[25] 数据集是图像领域的数据集，总计 194 个实例，其中特征个数为 19，标记个数为 7；Yeast[5] 数据集为生物领域的数据集，总计 2 417 个实例，其中特征个数为 103，标记个数为 14；Scene[3] 数据集为图像领域的数据集，总计 2 407 个实例，其中特征个数为 294，标记个数为 6。所有数据集的训练集和测试集都是预先划分好的，不需要进行手动划分。数据集的详细信息如表 1 所示。

表 1　数据集描述

Table 1　Description of data sets

Name	Domain	Instance	Attribute	Label	Train	Test	Density	Cardinality
Emotions[23]	Music	593	72	6	391	202	0.311	1.869
Birds[24]	Audio	645	260	19	322	323	0.053	1.014
Flags[25]	Image	194	19	7	129	65	0.485	3.392
Yeast[5]	Biology	2 417	103	14	1 500	917	0.303	4.237
Scene[3]	Image	2 407	294	6	1 211	1 196	0.179	1.074

评估分类模型的指标有以下 5 种：Average precision，Hamming loss，One error，Ranking loss 和 Coverage。这 5 种评价指标从多个角度对多标记分类模型进行了评价。本文全部实验的硬件环境是：2.6 GHz 的处理器，8 GB 的内存空间。

以下简要说明基学习器的生成过程，具体可见文献[26]。

首先，构造了多标记变精度邻域粗糙集模型，进而给出了该模型下的属性约简算法。由不同的邻域半径，精度可产生不同的子空间，基于不同的子空间可构造不同的学习器。

MDCCS 算法可通过计算总最小距离得到最佳簇数。其中，MDCCS 算法下的簇中心只能在已有基学习器集合中产生。KMCCS 算法基于 K-means 对所有基学习器进行聚类，最终可得簇划分结果。其中，KMCCS 算法下的簇中心可以依据已有基学习器的组合产生新的簇中心，但新的簇中心最终并不参与集成。

K-means 聚类算法中，距离的计算方式以及聚类簇数 K 的变化都会对最终结果产生影响。由于在多标记下的 5 种指标中只有平均精度与其他 4 种指标变化趋势相反，以下图示中仅涉及平均精度以及其他 4 种任意一种指标（此处随机选择排序损失）。以 Emotions 与 Birds 数据集为例，图 1，2 描述了平均精度与排序损失在 MDCCS 算法下的簇数 K 值的变化以及采用不同计算方式而产生的变化。图 3，4 描述了平均精度与排序损失在 KMCCS 算法下的簇数 K 值的变化以及采用不同计算方式而产生的变化。

分析 Emotions，Birds 数据集在 MDCCS 算法下的表现。如图 1 所示，4 种距离度量下的平均精度的变化趋势与排序损失的变化趋势均是相反的；欧氏距离与汉明距离的变化趋势相似。杰卡德距离的平均精度在簇数 K 取 [1，10] 内的值时波动较大。MDCCS 算法下的 4 种距离度量均在簇数 K 取 50 左右的值时平均精度达到最大，但杰卡德距

离在平均精度取最大值时对应的排序损失并不是其最小值，仅仅是簇数K在50附近的局部最小值，其排序损失的最小值对应的簇数K值在30附近。欧氏距离与汉明距离对应的平均精度较高值仅仅为簇数K=52时对应的值，而余弦距离对应的平均精度在[20,30],[30,40],[40,50]均有较高值，杰卡德距离对应的平均精度在[0,10],[20,30],[40,50]均有较高值。

如图2所示，4种距离度量下的平均精度的变化趋势与排序损失的变化趋势均是相反的；余弦距离与杰卡德距离的平均精度在簇数K取[5,10]内的值时波动较大，欧氏距离与汉明距离的平均精度在簇数K取[5,25]内的值时波动较大。欧氏距离下的平均精度在[5,10],[10,15],[20,25]均有较高值，余弦距离下的平均精度在[5,10],[30,35]均有较高值，汉明距离下的平均精度在[5,10],[20,25],[30,35]均有较高值，杰卡德距离下的平均精度仅在簇数K=6时达到较高值。

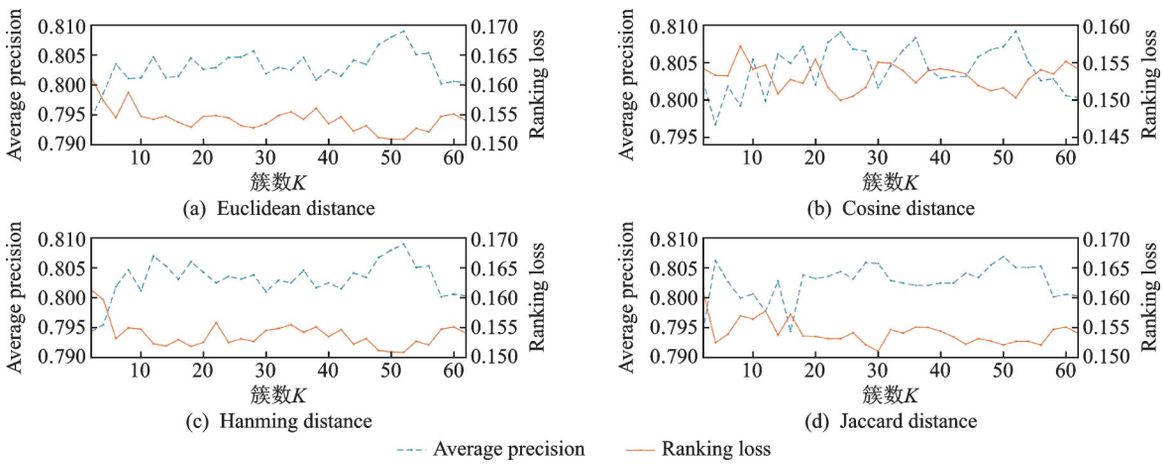

图1 MDCCS基于4种距离度量的平均精度与排序损失(Emotions)

Fig.1 Average precision and ranking loss with MDCCS based on four distance measures(Emotions)

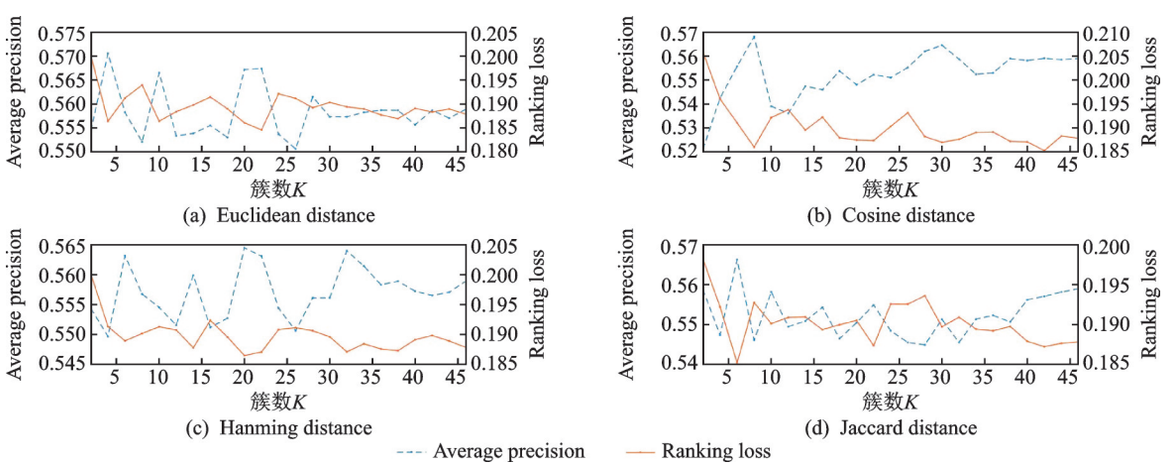

图2 MDCCS基于4种距离度量的平均精度与排序损失(Birds)

Fig.2 Average precision and ranking loss with MDCCS based on four distance measures(Birds)

分析Emotions,Birds数据集在KMCCS算法下的表现。如图3所示，4种距离度量下的平均精度的变化趋势与排序损失的变化趋势均是相反的。欧氏距离下的平均精度在[30,40],[40,50],[50,60]均有较高值；余弦距离下的平均精度在[30,40],[40,50],[50,60]均有较高值；汉明距离下的平均精度在[10,20],[30,40],[40,50]均有较高值；杰卡德距离下的平均精度在[20,30],[30,40],[40,50]均有较高值。

如图4所示，4种距离度量下的平均精度的变化趋势与排序损失的变化趋势均是相反的。欧氏距离下的平均精度在[5,10]有较高值；余弦距离下的平均精度在[30,35],[35,40]均有较高值；汉明距离下的平均精度在[5,10]有较高值；杰卡德距离下的平均精度在[35,40]有较高值。

在Emotions数据集下，MDCCS算法下的平均精度的变化趋势为先增后减；而KMCCS算法下的平均精度的变化趋势为先增后趋于平稳。在Birds数据集下，MDCCS算法下的平均精度的变化趋势为先增后减再增，最后趋于平稳，在簇数K逐渐增大的过程中波动较大；而在KMCCS算法下的平均精度的变化趋势为先增后减，最后趋于平稳，在簇

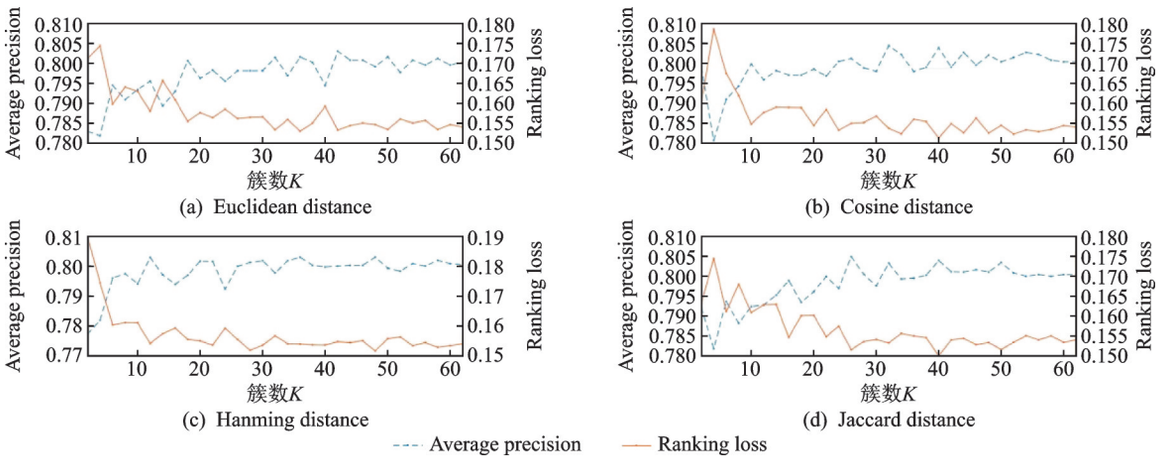

图 3　KMCCS 基于 4 种距离度量的平均精度与排序损失(Emotions)

Fig.3　Average precision and ranking loss with KMCCS based on four distance measures (Emotions)

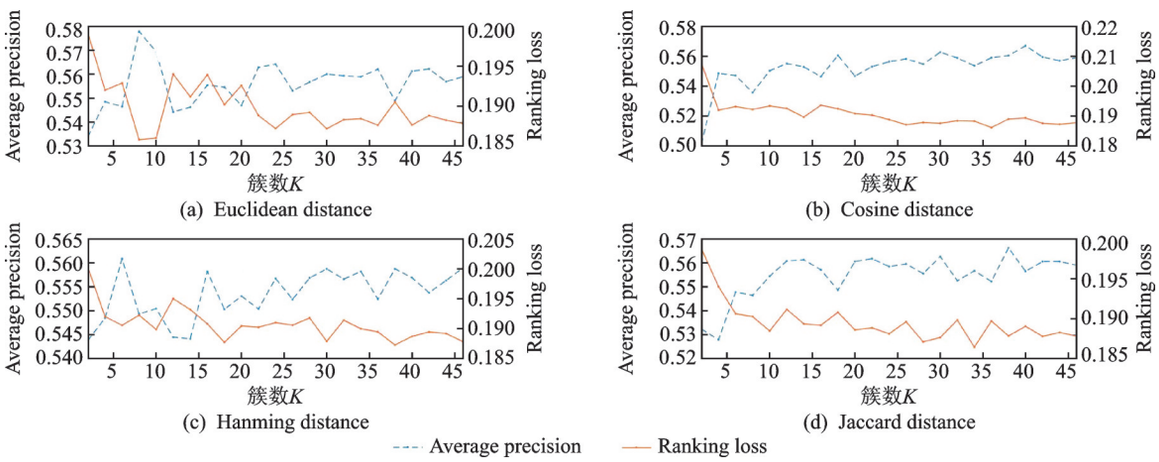

图 4　KMCCS 基于 4 种距离度量的平均精度与排序损失(Birds)

Fig.4　Average precision and ranking loss with KMCCS based on four distance measures (Birds)

数 K 逐渐增大的过程中波动较小。由图 1—4 可知，在不同的数据集下，簇数 K 计算方式的变化都可对评价指标的变化趋势产生影响。

表 2—11 给出了两种算法下得出的最优指标值与集成下得出的指标值之间的对比。其中 No1.euc，No1.cos，No1.ham，No1.jac，No2.euc，No2.cos，No2.ham，No2.jac 分别代表 MDCCS 算法下以及 KMCCS 算法下的欧氏距离、余弦相似度、汉明距离和杰卡德距离。由于 5 种指标的最大值或最小值对应的簇数并不相同，本文基于最大平均精度对集成前后的结果以及不同选择性集成算法得到的实验结果进行对比。

由表 2—11 可知，在 Emotions 数据集中，MDCCS 算法下的 4 种距离计算方式的平均精度高于集成后的平均精度；KMCCS 算法下的余弦相似度与杰卡德距离的平均精度高于集成后的平均精度，但欧氏距离与汉明距离低于集成后的平均精度；两种算法下的平均精度均略低于集成前的最高平均精度。

在 Birds 数据集中，MDCCS 算法下的欧氏距离，余弦相似度，杰卡德距离的平均精度高于集成后的平均精度，但汉明距离的平均精度低于集成后的平均精度；KMCCS 算法下的欧氏距离，余弦相似度，杰卡德距离的平均精度高于集成后的平均精度，但汉明距离的平均精度低于集成后的平均精度；且两种算法下的平均精度均高于集成前的最高平均精度。

表 2　MDCCS 基于 4 种距离度量的 5 种指标对比(Emotions)

Table 2　Comparison of five indicators of MDCCS based on four distance measures(Emotions)

集成	AP	CV	HL	OE	RL
集成前	0.818 8	1.811 9	0.192 2	0.247 5	0.152 2
集成后	0.803 9	1.802 0	0.189 8	0.297 0	0.151 7
No1.euc	0.808 9	1.816 8	0.195 5	0.277 2	0.150 9
No1.cos	0.809 1	1.811 9	0.195 5	0.277 2	0.150 3
No1.ham	0.808 9	1.816 8	0.195 5	0.277 2	0.150 3
No1.jac	0.806 8	1.821 8	0.197 2	0.282 2	0.152 1

表3 KMCCS基于4种距离度量的5种指标对比(Emotions)

Table 3 Comparison of five indicators of KMCCS based on four distance measures(Emotions)

集成	AP	CV	HL	OE	RL
集成前	0.818 8	1.811 9	0.192 2	0.247 5	0.152 2
集成后	0.803 9	1.802 0	0.189 8	0.297 0	0.151 7
No2.euc	0.803 0	1.797 0	0.191 4	0.306 9	0.153 3
No2.cos	0.804 4	1.806 9	0.192 2	0.297 0	0.153 7
No2.ham	0.802 9	1.797 0	0.192 2	0.311 9	0.151 7
No2.jac	0.804 8	1.802 0	0.198 0	0.302 0	0.151 6

表4 MDCCS基于4种距离度量的5种指标对比(Birds)

Table 4 Comparison of five indicators of MDCCS based on four distance measures(Birds)

集成	AP	CV	HL	OE	RL
集成前	0.559 0	2.380 8	0.046 0	0.494 2	0.193 6
集成后	0.565 1	2.294 1	0.048 9	0.505 8	0.187 0
No1.euc	0.570 5	2.291 0	0.048 6	0.494 2	0.186 4
No1.cos	0.567 9	2.291 0	0.048 4	0.511 6	0.186 0
No1.ham	0.564 4	2.294 1	0.048 9	0.517 4	0.186 5
No1.jac	0.566 2	2.278 6	0.047 7	0.511 6	0.185 2

表5 KMCCS基于4种距离度量的5种指标对比(Birds)

Table 5 Comparison of five indicators of KMCCS based on four distance measures(Birds)

集成	AP	CV	HL	OE	RL
集成前	0.559 0	2.380 8	0.046 0	0.494 2	0.193 6
集成后	0.565 1	2.294 1	0.048 9	0.505 8	0.187 0
No2.euc	0.577 7	2.281 7	0.047 4	0.482 6	0.185 8
No2.cos	0.566 8	2.322 0	0.047 9	0.511 6	0.189 4
No2.ham	0.560 8	2.306 5	0.048 7	0.511 6	0.190 6
No2.jac	0.566 1	2.315 8	0.047 9	0.511 6	0.187 8

表6 MDCCS基于4种距离度量的5种指标对比(Yeast)

Table 6 Comparison of five indicators of MDCCS based on four distance measures(Yeast)

集成	AP	CV	HL	OE	RL
集成前	0.761 9	6.191 2	0.195 2	0.224 6	0.167 7
集成后	0.759 9	6.308 6	0.196 8	0.237 7	0.168 1
No1.euc	0.760 8	6.291 2	0.195 7	0.239 9	0.168 0
No1.cos	0.760 8	6.291 2	0.195 7	0.239 9	0.168 0
No1.ham	0.761 3	6.281 4	0.195 4	0.235 6	0.167 7
No1.jac	0.760 8	6.291 2	0.195 7	0.239 9	0.168 0

在Yeast数据集中,MDCCS算法下的4种距离计算方式的平均精度高于集成后的平均精度;KMCCS算法下的4种距离计算方式的平均精度高于集成后的平均精度;但MDCCS算法下的平均精度均低于集成前的平均精度,KMCCS算法下的欧氏距离,汉明距离均高于集成前的平均精度。

在Flags数据集中,MDCCS算法下的4种距离计算方式的平均精度高于集成后的平均精度;KMCCS算法下的欧氏距离、汉明距离和杰卡德距离的平均精度高于集成后的平均精度,但余弦相似度的平均精度低于集成后的平均精度;且MDCCS算法下的欧氏距离、汉明距离和杰卡德距离略高于集成前的最高平均精度。

表7 KMCCS基于4种距离度量的5种指标对比(Yeast)

Table 7 Comparison of five indicators of KMCCS based on four distance measures(Yeast)

集成	AP	CV	HL	OE	RL
集成前	0.761 9	6.191 2	0.195 2	0.224 6	0.167 7
集成后	0.759 9	6.308 6	0.196 8	0.237 7	0.168 1
No2.euc	0.762 6	6.265 0	0.194 3	0.235 6	0.166 1
No2.cos	0.760 6	6.274 8	0.195 6	0.237 7	0.167 4
No2.ham	0.763 0	6.275 9	0.197 0	0.234 5	0.166 4
No2.jac	0.761 0	6.266 1	0.195 3	0.237 7	0.166 9

表8 MDCCS基于4种距离度量的5种指标对比(Scene)

Table 8 Comparison of five indicators of MDCCS based on four distance measures(Scene)

集成	AP	CV	HL	OE	RL
集成前	0.856 2	0.549 3	0.092 1	0.231 6	0.088 9
集成后	0.853 7	0.546 0	0.099 6	0.240 0	0.088 4
No1.euc	0.855 0	0.546 8	0.099 5	0.236 6	0.088 5
No1.cos	0.855 6	0.538 5	0.099 4	0.235 8	0.086 8
No1.ham	0.855 0	0.541 8	0.099 4	0.238 3	0.086 7
No1.jac	0.856 3	0.539 3	0.099 1	0.235 8	0.087 5

表9 KMCCS基于4种距离度量的5种指标对比(Scene)

Table 9 Comparison of five indicators of KMCCS based on four distance measures(Scene)

集成	AP	CV	HL	OE	RL
集成前	0.856 2	0.549 3	0.092 1	0.231 6	0.088 9
集成后	0.853 7	0.546 0	0.099 6	0.240 0	0.088 4
No2.euc	null	null	null	null	null
No2.cos	null	null	null	null	null
No2.ham	0.856 1	0.521 7	0.098 1	0.239 1	0.083 9
No2.jac	0.856 5	0.527 6	0.099 2	0.236 6	0.084 8

表10 MDCCS基于4种距离度量的5种指标对比(Flags)

Table 10 Comparison of five indicators of MDCCS based on four distance measures(Flags)

集成	AP	CV	HL	OE	RL
集成前	0.822 7	3.661 5	0.268 1	0.169 2	0.206 9
集成后	0.821 2	3.600 0	0.274 7	0.215 4	0.204 4
No1.euc	0.824 1	3.553 8	0.272 5	0.215 4	0.196 4
No1.cos	0.822 5	3.600 0	0.259 3	0.215 4	0.201 0
No1.ham	0.823 5	3.584 6	0.274 7	0.215 4	0.198 2
No1.jac	0.822 9	3.584 6	0.270 3	0.215 4	0.198 5

在Scene数据集中,MDCCS算法下的4种距

表 11　KMCCS基于4种距离度量的5种指标对比(Flags)

Table 11　Comparison of five indicators of KMCCS based on four distance measures(Flags)

集成	AP	CV	HL	OE	RL
集成前	0.822 7	3.661 5	0.268 1	0.169 2	0.206 9
集成后	0.821 2	3.600 0	0.274 7	0.215 4	0.204 4
No2.euc	0.822 5	3.615 4	0.268 1	0.215 4	0.200 8
No2.cos	0.821 1	3.615 4	0.272 5	0.215 4	0.202 1
No2.ham	0.821 9	3.615 4	0.270 3	0.215 4	0.202 1
No2.jac	0.821 9	3.600 0	0.265 9	0.215 4	0.200 3

离计算方式的平均精度高于集成后的平均精度；KMCCS算法下的汉明距离、杰卡德距离的平均精度高于集成后的平均精度；但两种算法下的杰卡德距离略高于集成前的最高平均精度，而KMCCS算法下欧氏距离与余弦相似度对应的指标值标记为null的原因是K-means聚类算法中出现空簇导致算法提前停止。

由于在Scene数据集下使用本文所提出的KMCCS算法出现空簇现象，即某一次计算过程中某簇不包含任何学习器的现象，本文提出了两种解决方法尝试解决该问题。

第1种解决方法为令产生空簇的质心不变，继续使用该质心参与下一次迭代；第2种解决方法为在当前迭代结果的簇内总距离最大的簇中随机选取一个作为新的质心作为候选补充到空簇中。其中，1-euc，1-cos分别代表第1种解决方法下的欧氏距离与余弦相似度；2-euc，2-cos分别代表第2种解决方法下的欧氏距离与余弦相似度。实验结果见表12。

表 12　两种解决空簇方法的对比

Table 12　Comparison of two solutions to empty cluster

集成	AP	CV	HL	OE	RL
1-euc	null	null	null	null	null
1-cos	0.856 0	0.526 8	0.098 9	0.239 1	0.084 9
2-euc	0.856 2	0.522 6	0.099 8	0.239 1	0.084 0
2-cos	0.856 6	0.528 4	0.098 9	0.236 6	0.085 2

由表12可知，两种解决方法均可用于解决空簇问题，在第2种解决方法下的欧氏距离的平均精度与集成前的最高平均精度相同；余弦相似度的平均精度高于集成前的最高平均精度。

4　结　　论

本文提出了多标记学习框架下的两种基于聚类的选择性集成算法。一种是基于最小总距离的簇中心选择算法，另一种是基于K-means提出的簇中心选择算法，对这两种算法进行了一系列对比实验。另外，由于KMCCS算法在Scene数据集下的计算过程中产生了空簇，给出了两种方法解决空簇问题。

在未来的工作中，可以尝试考虑选择其他的多标记选择性集成，例如基于优化的多标记选择性集成或基于排序的多标记选择性集成等。

参 考 文 献：

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

[2] CLARE A, KING R D. Knowledge discovery in multi-label phenotype data[J]. Lecture Notes in Computer Science, 2001, 2168: 42-53.

[3] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[4] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge: MIT Press, 2001: 681-687.

[5] ZHANG Minling, ZHOU Zhihua. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[6] ZHANG Minling. Ml-RBF: RBF neural networks for multi-label learning[J]. Neural Processing Letters, 2009, 29(2): 61-74.

[7] XIAO Yawen, WU Jun, LIN Zongli, et al. A deep learning-based multi-model ensemble method for cancer prediction[J]. Computer Methods and Programs in Biomedicine, 2018, 153: 1-9.

[8] ELGHAZEL H, AUSSEM A, GHARROUDI O, et al. Ensemble multi-label text categorization based on rotation forest and latent semantic indexing[J]. Expert Systems with Applications, 2016, 57: 1-11.

[9] 艾科，马国帅，杨凯凯，等. 一种基于集成学习的科研合作者潜力预测分类方法[J]. 计算机研究与发展，2019，56(7)：1383-1395.
AI Ke, MA Guoshuai, YANG Kaikai, et al. A classification method of scientific collaborator potential prediction based on ensemble learning[J]. Journal of Computer Research and Development, 2019, 56(7): 1383-1395.

[10] ZHANG Xiaofei, OUYANG Le, YANG Shuo, et al. EnImpute: Iimputing dropout events in single-cell RNA-sequencing data via ensemble learning[J]. Bioinformatics, 2019, 35(22): 4827-4829.

[11] MA Zhuo, LIU Yang, LIU Ximeng, et al. Lightweight privacy-preserving ensemble classification for

face recognition[J]. IEEE Internet of Things Journal, 2019, 6(3): 5778-5790.

[12] 冯代高, 张友俊. 改进随机子空间LDA结合多补丁集成学习的鲁棒人脸识别算法[J]. 计算机应用研究, 2019, 36(8): 2556-2560.
FENG Daigao, ZHANG Youjun. Robust face recognition algorithm based on multiple patch integration learning and improved random subspace LDA[J]. Application Research of Computers, 2019, 36(8): 2556-2560.

[13] ALZUBI J A, BHARATHIKANNAN B, TANWAR S, et al. Boosted neural network ensemble classification for lung cancer disease diagnosis[J]. Applied Soft Computing, 2019, 80: 579-591.

[14] 魏秀参, 慕鑫, 杨杨. 二次集成学习在医疗数据挖掘中的应用[J]. 计算机科学与探索, 2014, 8(9): 1113-1119.
WEI Xiushen, MU Xin, YANG Yang. An application in medical data mining based on twice ensemble learning[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(9): 1113-1119.

[15] ZHOU Zhihua, WU Jianxin, TANG Wei. Ensembling neural networks: Many could be better than all [J]. Artificial Intelligence, 2002, 137(1): 239-263.

[16] LIN Chen, CHEN Wenqiang, QIU Cheng, et al. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy[J]. Neurocomputing, 2014, 123: 424-435.

[17] CROUX C, JOOSSENS K, LEMMENS A. Trimmed bagging[J]. Computational Statistics & Data Analysis, 2007, 52(1): 362-368.

[18] ZHOU Zhihua, WU Jianxin, TANG Wei, et al. Combining regression estimators: GA-based selective neural network ensemble[J]. International Journal of Computational Intelligence and Applications, 2001, 1(4): 341-356.

[19] XING Hongjie, WANG Xizhao. Selective ensemble of SVDDs with Renyi entropy based diversity measure [J]. Pattern Recognition, 2017, 61: 185-196.

[20] ZHANG Yuanjian, MIAO Duoqian, ZHANG Zhifei, et al. A three-way selective ensemble model for multi-label classification[J]. International Journal of Approximate Reasoning, 2018, 103: 394-413.

[21] WU Qingyao, TAN Mingkui, SONG Hengjie, et al. ML-FOREST: A multi-label tree ensemble method for multi-label classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2665-2680.

[22] JAIN A K. Data clustering: 50 years beyond *K*-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[23] TROHIDIS K, TSOUMAKAS G, KALLIRIS G, et al. Multilabel classification of music into emotion [C]//Proceedings of the 9th Int Society for Music Information Retrieval. Philadelphia: ISMIR, 2008: 325-330.

[24] BRIGGS F, HUANG Y, RAICH R, et al. The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment[C]//Proceedings of 2013 IEEE Int Workshop on Machine Learning for Signal Processing. Los Alamitos, CA: IEEE, 2013: 22-25.

[25] EDUARDO C G, PLASTINO A, FREITAS A A. A genetic algorithm for optimizing the label ordering in multi-label classifier chains[C]//Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon, VA, USA: IEEE, 2013: 469-476.

[26] 张佳欢, 李磊军, 李美争, 等. 基于变精度邻域粗糙集的多标记子空间研究[J]. 南京理工大学学报(自然科学版), 2019, 43(4): 414-422.
ZHANG Jiahuan, LI Leijun, LI Meizheng, et al. Research on multi-label subspace based on variable precision neighborhood rough sets[J]. Journal of Nanjing University of Science and Technology, 2019, 43(4): 414-422.

(编辑:孙静)