



文章编号:1001-9081(2019)09-2499-06

DOI:10.11772/j.issn.1001-9081.2019020763

基于最远总距离采样的代价敏感主动学习

任杰¹, 阎帆^{1*}, 汪敏²

(1. 西南石油大学 计算机科学学院, 成都 610500; 2. 西南石油大学 电气信息学院, 成都 610500)

(*通信作者电子邮箱 minfanphd@163.com)

摘要:主动学习旨在通过人机交互减少专家标注,代价敏感主动学习则致力于平衡标注与误分类代价。基于三支决策(3WD)和标签均匀分布(LUD)模型,提出一种基于最远总距离采样的代价敏感主动学习算法(CAFS)。首先,设计了最远总距离采样策略,以查询代表性样本的标签;其次,利用了LUD模型和代价函数,计算期望采样数目;最后,使用了k-Means聚类技术分裂已获得不同标签的块。CAFS算法利用三支决策思想迭代地进行标签查询、实例预测和块分裂,直至处理完所有实例。学习过程在代价最小化目标的控制下进行。在9个公开数据上比较,CAFS比11个主流的算法具有更低的平均代价。

关键词:主动学习; k-Means聚类; 标签均匀分布; 三支决策

中图分类号: TP181 文献标志码:A

Cost-sensitive active learning through farthest distance sum sampling

REN Jie¹, MIN Fan^{1*}, WANG Min²

(1. School of Computer Science, Southwest Petroleum University, Chengdu Sichuan 610500, China;

2. School of Electrical Engineering and Information, Southwest Petroleum University, Chengdu Sichuan 610500, China)

Abstract: Active learning aims to reduce expert labeling through man-machine interaction, while cost-sensitive active learning focuses on balancing labeling and misclassification costs. Based on Three-Way Decision (3WD) methodology and Label Uniform Distribution (LUD) model, a Cost-sensitive Active learning through the Farthest distance sum Sampling (CAFS) algorithm was proposed. Firstly, the farthest total distance sampling strategy was designed to query the labels of representative samples. Secondly, LUD model and cost function were used to calculate the expected sampling number. Finally, k-Means algorithm was employed to split blocks obtained different labels. In CAFS, 3WD methodology was adopted in the iterative process of label query, instance prediction, and block splitting, until all instances were processed. The learning process was controlled by the cost minimization objective. Results on 9 public datasets show that CAFS has lower average cost compared with 11 mainstream algorithms.

Key words: active learning; k-Means clustering; label uniform distribution; Three-Way Decision (3WD)

0 引言

主动学习^[1]是半监督学习^[2]的一种方式,旨在通过人机交互减少专家标注的工作量。常用方法大致分为两类:基于聚类的方法选择具有代表性的对象,基于委员会的方法^[3]选择不确定性高的对象。Cohn等^[4]提出了一种基于高斯模型和局部加权回归模型的主动学习算法,应用模型以及回归使主动学习所需的训练样本急剧减少。Wang等^[5]提出了基于密度峰值聚类的主动学习算法,在相同的训练样本基础上使得算法的分类精度进一步提高。目前主动学习已广泛应用于文本分类^[6]、信息提取^[7]、图像分类^[8]、语音识别^[9]等领域。

代价敏感主动学习^[10]致力于平衡标注与误分类代价。教师代价是专家标注样本标签的代价,误分类代价是指将样本错误分类的代价。该问题比经典的主动学习更有实际意义,也更具一般性。Min等^[11]利用k最近邻(k-Nearest Neighbors, kNN)将总体根据代价分成3个部分,提出了基于kNN的三分代价敏感主动学习算法,该算法重复三分区过程

从而减少了总代价;但该算法并未考虑块内采样数目。Wu等^[12]建立了标签均匀分布模型,在代价的基础上利用标签均匀分布(Label Uniform Distribution, LUD)模型计算每块内最优的采样数目,进一步降低了代价;但其采样策略没有考虑样本点的信息量,使得代价依然有可优化的空间。

本文提出一种基于最远总距离采样的代价敏感主动学习算法(Cost-sensitive Active learning through the Farthest distance sum Sampling, CAFS)。该算法有如下特点:

1) 利用三支决策(Three-Way Decision, 3WD)的思想,使学习过程更加完善。算法迭代地进行标签查询、实例预测和块分裂,直至处理完所有实例。方案在查询过程中进行分类,不需要引入其他的分类器。

2) 提出了最远总距离策略以获得需查询标签的样本。针对随机采样采样的不足,该策略综合考虑了某块内已查询的所有样本和信息量,可获得更具代表性样本。

3) 采用LUD模型计算块内需要查询的样本数,并设置阈值,对过小的块进行总体查询,使得采样数目在此情况下达到

收稿日期:2019-03-22;修回日期:2019-05-06;录用日期:2019-05-29。

基金项目:四川省青年科技创新团队专项(2019JDTD0017);四川省应用基础研究项目(2019JDTD0017)。

作者简介:任杰(1996—),男,山西忻州人,硕士研究生,主要研究方向:主动学习; 阎帆(1973—),男,重庆人,教授,博士,CCF会员,主要研究方向:粒计算、推荐系统、主动学习; 汪敏(1980—),女,湖南邵阳人,副教授,硕士,CCF会员,主要研究方向:数据挖掘、主动学习。



最优。该模型对不同的数据集有较好的适用性。

4) 采用了高效的 *k*-Means 聚类算法。该算法使用距离函数表达对象的相似性,与最远总距离采样策略配合可以获得很好效果。

本文在 9 个数据集上与 11 个主流算法进行了比较,结果表明,CAFS 算法在平均代价方面优于对比算法。

1 相关工作

1.1 三支决策

三支决策(3WD)^[13]是一种符合人类认知的决策模式。它是实现二支决策的一个中间步骤,在实际决策的过程中,对于具有充分把握接受或拒绝的事物能够立即作出快速的判断,对于那些不能立即作出决策的事件,则进行延迟决策。三支决策是一种包含三个部分或三个操作的分治方法,也是决策理论粗糙集的延伸。

很多理论和应用使用了三支决策的方法及思想。其中三支形式概念分析和三支认知计算衍生出了概念学习和多粒度认识操作。通过决策粗糙集理论和属性约简方法将三支决策理论粗糙集与代价敏感相结合^[14],在样本上得出最优测试属性,并依据最优测试属性在测试集上计算,使得分类结果具有最小误分类代价和测试代价。基于三支决策的多粒度粗糙集理论^[15]通过分析三支决策与概率粗糙集、决策粗糙集间的关系以及在属性约简的相关知识,给出了在医学、工程方向的应用和三支决策未来的发展方向。三支邻域粗糙集模型^[16]根据错误率和多粒度构建不同的邻域系统,证明了可变精度粗糙集和多粒度粗糙集是邻域系统粗糙集模型的特例。

1.2 代价敏感主动学习

代价敏感主动学习在主动学习的基础上,考虑了在学习过程中的代价敏感性,为不同的类别提供了不同的代价权重以及教师代价,在代价函数的约束下进行学习。

由于代价敏感学习更具有实际意义,从而受到很多学者的关注,如文献[10]中引入了代价敏感主动学习,并提出在未标记数据下的分类概率和基于分类概率的抽样和决策。Settles 等^[17]分析了 4 个真实的文字和图像领域的教师代价,给出了某些具体领域的教师代价的特征。Liu 等^[18]将联系教师代价与距离,使代价敏感主动学习在空间数据上展开。Zhao 等^[19]通过优化两种代价处理不平衡 URL 检测任务的问题,使代价敏感主动学习在 URL 检测问题上优于一般检测学习算法。Chen 等^[20]提出了最大预期代价和代价加权边际最小策略,使多类代价敏感主动学习表现更加突出。Demir 等^[21]通过在遥感图像分类中,使用成本函构建教师代价利用了遥感图像的特性,使代价的定义更为全面。Huang 等^[22]通过非度量多位缩放将代价信息嵌入到特殊隐藏空间中的距离中,从隐藏空间的距离定义样本的不确定性,使学习过程选择更有效的样本。

1.3 标签均匀分布模型

目前,数据集中大量标签未知是造成多种学习任务结果不理想的重要原因之一,主动学习算法正是此类问题的合理解决方案。对于大量标签未知的数据,我们很迫切地需要知道数据的结构以及分布,所以很容易基于现实模型或者简单的理论分析来假设一种分布模型,应用数据本身的结构在满足任务目标的前提下降低学习过程中的代价。

基于最远总距离采样的代价敏感主动学习 CAFS 算法应

用简单的均匀分布统计模型,利用概率和均值估计二分类数据中的正反例的个数。同时为了减少总教师代价,在均匀分布的基础上,利用期望数目和代价函数计算最优采样数目。

CAFS 算法采用标签均匀分布模型,即在总体分布未知的情况下,假设二分类总体中抽到正反例的概率相同。其概率如下:

$$p(R^* = i) = \frac{1}{n+1}; \quad \forall 0 \leq i \leq n \quad (1)$$

在标签均匀分布模型中,如果在总体 X 中随机选取 R 个正例和 B 个反例,那么在总体中有 R^* 个正例的概率则为:

$$\bar{b}(R^* | R, B; n) = \frac{\sum_{i=R}^{n-R} i A_i^R A_{n-i}^B}{\sum_{i=R}^{n-R} A_i^R A_{n-i}^B} \quad (2)$$

在上述假设以及概率公式成立的情况下,正反例在总体 X 中期望的数目为:

$$\bar{b}(n, R, B) = \bar{r}(n, B, R) = \frac{\sum_{i=R}^{n-R} i A_i^R A_{n-i}^B}{n \sum_{i=R}^{n-R} A_i^R A_{n-i}^B} \quad (3)$$

当在连续抽出正例或反例时候,出现另一个对立的实例对于期望的影响很大,有如下公式成立:

$$\bar{r}(n, R, 0) > \bar{r}(n, 2R-1, 1) \quad (4)$$

2 代价敏感主动学习问题描述

为介绍 CAFS 算法,表 1 列出了本文使用的符号以及含义。

表 1 符号以及含义

Tab. 1 Symbols and meanings

变量符号	含义
U	样本总体
C	条件属性集
D	决策属性
V_a	属性 a 的值域, $a \in C \cup D$
V	V_a 的集合
I	信息函数
m	误分类代价
t	教师代价
x_i	U 中第 i 个实例
y_i	x_i 的实际标签
l_i	x_i 的预测标签
SL_i	选择的代表点 i
N	U 中的实例个数
X	U 的一个子集
n	X 中的实例个数
R	X 中已经被查询过的正样本个数
B	X 中已经被查询过的负样本个数
A_i^j	从 i 个实例中取 j 个做排列数 A
$P(R^* R, B; n)$	由 X 中抽出 R 个正例, B 个反例, 则 X 中恰好含有 R^* 个正例的概率
$\bar{r}(n, R, B)$	X 中正实例的期望比例
$\bar{b}(n, R, B)$	X 中负实例的期望比例
$\sigma(n, R, B)$	正实例比例的标准差
f	X 中期望采样数目
s	X 中最优采样数目

2.1 数据模型

CAFS 算法使用如下数据模型。



定义 1 教师误分类代价敏感决策系统
(Teacher-and-Misclassification-Cost-sensitive Decision System, TMC-DS),是七元组:

$$S = (U, C, d, V, I, m, t) \quad (5)$$

其中: U 是有限的实例集合, C 是条件属性的集合, d 是代价属性, $V = \bigcup_{a \in C \cup \{d\}} V_a$, V_a 是属性 a 的属性值, $I: U \times (C \cup \{d\}) \rightarrow V$ 是信息函数, $m: V_d \times V_d \rightarrow R^+ \cup \{0\}$ 是误分类代价函数, $t \in R^+ \cup \{0\}$ 是教师代价。

2.2 问题定义

问题 1 代价敏感主动学习。

输入:一个代价敏感决策系统七元组 TMC-DS;

输出:专家查询的实例集合 U_t ,预测标签 l_{U-U_t} 。

优化目标: $\min cost = \left(t | U_t | + \sum_{i=1}^{|U|} m(l_i, y_i) \right) / |U|$

输入的是不含标记的代价敏感决策系统 TMC-DS。输出包含两个部分:其一是实例子集 U_t ,其中的标签是查询或者由专家给出;其二是剩余实例的预测标签 l_{U-U_t} 。

优化目标是通过减少教师代价和误分类代价使平均代价达到最小,其中 $t \times |U_t|$ 是总教师代价, $\sum_{i=1}^{|U|} m(l_i, y_i)$ 是总的误分类代价。其中教师代价和误分类代价是在获得 U_t 之后计算得到的, U_t 并不是用户指定的。而随着 U_t 大小的增加,教师代价呈线性增长,误分类代价可能会减少,本文的 CAFS 算法找到了一个教师代价与误分类代价的相对平衡点。

3 CAFS 算法

本章将详细介绍 CAFS 算法的执行过程,其中包括 CAFS 算法总体流程、根据 LUD 模型以及代价函数计算出最优采样数目的 lookup 方法、根据最远总距离采样策略利用 k-Means 聚类对块进行分裂并迭代学习的 splitAndLearn 方法。

3.1 算法框架

基于最远总距离采样的代价敏感主动学习 CAFS 的算法框架如算法 1 所示,其中第 2) 行是为了在块中寻找最远总距离的代表点,之后的步骤会确定当前块是否需要分块迭代学习。

算法 1 基于最远总距离采样的代价敏感主动学习算法 (CAFS)。

输入:样本总体 U ,算法 2(lookup) 最优采样数目 s ;
输出:预测标签集合 l_{U-U_t} 。

- 1) for($x_i \in U \&& (R \text{ or } B) < s$)
- 2) $SL_f \leftarrow \text{findFarthest}(U_t);$
/* 结合已经查询的实例结合中寻找最远距离点 */
- 3) if ($y_{SL_f} = y_0$)
/* 判断最远的代表样本点是否与初始样本点的标签相同 */
- 4) $U_t \leftarrow SL_f$
- 5) continue
- 6) else
- 7) splitAndLearn /* 分裂迭代学习算法 3 */
- 8) end if
- 9) end for
- 10) return l_{U-U_t}

算法 2 是 CAFS 算法中根据 LUD 模型计算要查询标签个

数的 lookup 方法,其中 f 是根据 LUD 模型以及代价函数所确定的正反例期望查询数目,如式(6)所示:

$$f = \begin{cases} m(-, +)N(1 - \bar{r}(N, R, 0)) + tR, & \text{已标记实例为正例} \\ m(+, -)N(1 - \bar{b}(N, 0, B)) + tB, & \text{已标记实例为反例} \end{cases} \quad (6)$$

算法 2 最优标签查询数目计算算法(lookup)。

输入:数据块的大小 n ,第一个抽出的样本标签 y_0 ;

输出:最优采样数 s 。

- 1) for($x_i \in X$)
- 2) $SL_f \leftarrow \text{bought}_i$ /* 记录已购买的标签 */
- 3) $i(r^*, b^*) \leftarrow f$ /* 根据式(6)计算期望查询数目 */
- 4) end for
- 5) $i^* \leftarrow \text{lookup}(y_0)$
- 6) $s \leftarrow (i^* - SL.length)$
- 7) return s

算法 3 介绍块分裂条件以及如何迭代学习的过程。在选取最远总距离代表点后,需要得知该代表点与之前查询的块标签是否一致。如果一致,继续利用最远距离采样策略采样直至达到最优采样数 s ,否则利用 k-Means 聚类算法分裂该块并迭代学习的过程。

算法 3 块分裂迭代学习算法(splitAndLearn)。

输入:数据块 X ;

输出:数据块的 X 的预测标签合集 $l_{i \in X}$ 。

- 1) if ($SL.length < s$)
- 2) $SL_{\text{new}} \leftarrow \text{findFarthest}(U_t)$
- 3) if ($y_{SL_{\text{new}}} \neq y_{x_0}$)
- 4) $X_1, X_2 \leftarrow \text{kMeansCluter}(X)$
- 5) $l_{x_1} \leftarrow \text{CAFS}(X_1)$
- 6) $l_{x_2} \leftarrow \text{CAFS}(X_2)$
- 7) end if
- 8) else
- 9) end if
- 10) return $l_{i \in X}$

3.2 CAFS 时间复杂度分析

基于最远总距离采样的代价敏感主动学习算法(CAFS)的时间复杂度如表 2 所示。

表 2 CAFS 时间复杂度
Tab. 2 Time complexity of CAFS

步骤描述	时间复杂度
计算最优代表点个数	$O(n \log n)$
寻找代表点	$\Theta(n \log n)$
分裂数据块	$O(n)$
预测其他实例标签	$\Theta(n)$
扫描数据	$\Theta(n)$

在实际算法过程中,时间复杂度会随着数据集变化而变化。在最优情况下,数据集中的实例个数趋于无穷时且为同一标签时,时间复杂度为 $O(n \log n)$ 。在最坏的情况下,且不同标签数据分布极为密集时,算法需要递归循环 $\log n$ 次。即:

$$\log n \times (O(n \log n) + \Theta(n \log n) + O(n)) = O(n^2)$$

4 CAFS 运行实例

CAFS 算法首先扫描块内已经标注的实例,查看该块是否



需要分裂,之后计算需要查询的个数,以最远总距离采样策略选取代表点并查询标签,在满足最优采样数目 s 后预测其他未标记数据。为了更好地展示 CAFS 算法的学习过程,以下将利用小型的数据集描述 CAFS 算法的学习过程。

如图 1 运行实例,首先,对数据进行初始化图 1(a),并记录数据中的第一个实例,然后如图 1(b)采用最远总距离样本采样策略选取代表性样本点查询。如图 1(c),很明显两个实例的标签不同,此时利用 k -Means 聚类对数据块进行分裂处理。对分裂后的数据块采样同样的策略迭代学习,如图 1(d) (e)。由于本次运行实例采用极具代表性的数据,所以在数据块的最优查询数目 s 的前提下两个块中的样本标签相同,根据 CAFS 算法的策略,如图 1(f)会对剩余的实例进行预测,即完成本次过程。

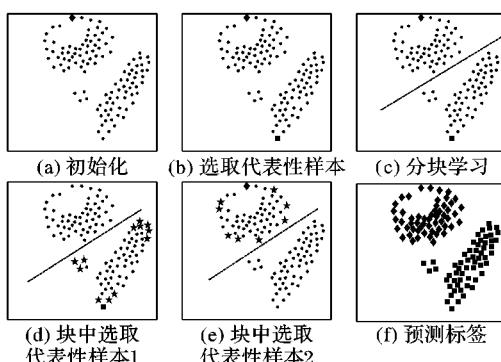


图 1 运行实例示意图
Fig. 1 Running example

5 实验与结果分析

实验运行在 64 位 16 GB RAM 的 Windows10 的个人电脑上,其中处理器为 i7-7700HQ 2.80 GHz,并利用 Java 在 Eclipse 上实现。

5.1 实验数据集

实验数据集来源于 UCI 机器学习仓库和 IDA 基准仓库,表 3 列出了数据集的基本信息,这些数据集一部分是人造数据集,大部分来源于现实生活,涵盖了生物学、金融学、计算机、通信、植物学、医疗和质谱分析等领域。

实验选取 11 个相关的算法进行了对比,并根据 CAFS 算

法特点分成了三组进行了相关实验:1)与同类的代价敏感学习算法进行对比;2)与代价敏感的主动学习算法进行对比;3)代价敏感学习与非代价敏感学习算法对比,而且为了将非代价敏感学习与代价敏感学习进行代价方面的对比,利用实验中的代价误分类代价设置,将非代价敏感学习的结果统一成代价进行比较。

表 3 数据集信息

Tab. 3 Dataset information

编号	名称	实例个数	实例维度	来源
1	Allmal	72	7 129	生物学
2	Arcene	200	1 000	质谱
3	Banana	5 300	2	植物学
4	Credit6000	5 987	65	金融学
5	Heart	270	13	医疗
6	Ionosphere	351	34	物理学
7	Madelon	2 600	500	人造
8	Sonar	208	60	通信
9	Spambase	4 207	57	计算机

5.2 实验代价设置

m 表示误分类代价矩阵, $m(+, -) = 4$ 表示将正例预测成反例的代价为 4, $m(-, +) = 2$ 表示将反例预测成正例的代价为 2。另外设置 $t = 1$ 是指查询一个实例的教师代价是 1。实验中的平均代价计算公式则为:

$$AverageCost = (M_1 m(+, -) + M_2 m(-, +) + tT)/n \quad (7)$$

其中, M_1 实验结果中将正例预测成反例的个数, M_2 为将反例预测成正例的个数, T 为向专家查询实例的个数。

5.3 与代价敏感学习算法的对比实验

本节将 CAFS 算法与代价敏感逻辑回归算法 (Cost Sensitive Logistic Regression algorithm, CSLR)^[23]、代价敏感决策树算法 (Cost Sensitive Decision Tree algorithm, CSDT)^[24] 和代价敏感随机森林算法 (Cost Sensitive Random Forest algorithm, CSRF)^[25] 在 9 个公开数据集上进行了对比,并以平均代价(根据式(7)计算)为唯一参照,结果如表 4 所示。其中“—”表示 CSLR 在 Arcene 数据集上运行超过 5 h 也没有产生结果;平均排名则指算法在所有数据集上表现排名的均值。从表 4 中看出,CAFS 的平均代价相对于 CSLR、CSDT、CSRF 分别降低了 56%、27%、32%。

表 4 CAFS 算法与其他代价敏感学习算法在不同数据集上的平均代价

Tab. 4 Comparison of average cost of CAFS algorithm and other cost-sensitive learning algorithms on different datasets

算法	Allmal	Arcene	Banana	Credit6000	Heart	Ionosphere	Madelon	Sonar	Spambase	平均排名
CSLR	1.1944	—	1.8377	1.6359	1.2407	0.8843	1.7523	1.3644	1.6463	3.8889
CSDT	1.1389	1.0350	0.7163	0.5927	0.9593	0.6074	1.3077	1.0644	0.4838	2.6667
CSRF	1.4167	1.0150	1.5559	0.5100	0.8896	0.5436	1.1992	1.0375	0.3426	2.2222
CAFS	0.7500	0.7050	0.4226	0.4808	0.6259	0.5128	0.8485	0.6827	0.6881	1.2222

5.4 与其他代价敏感主动学习算法的对比

本节实验选取了 5 个代价敏感主动学习算法进行比较。其中:ALCE(Active Learning Embed Cost algorithm)^[25] 为代价嵌入主动学习算法, CWMM (Cost Weight Minimum Margin algorithm) 为代价权重最小边缘算法, MEC(Maximum Expected Cost algorithm) 为最大期望代价算法, TALK (Tri-partition Active Learning through K-nearest neighbors algorithm) 为基于 k

近邻的三支决策主动学习算法, CADU (Cost-sensitive Active learning algorithm with a label Uniform Distribution model) 为基于密度聚类的代价敏感主动学习算法。

对 ALCE、CWMM 和 MEC 进行了 5 次重复实验,以保证实验结果的准确性;而且由于数据顺序不影响 TALK、CADU 和 CAFS 的结果,即实验的结果稳定,所以只进行 1 次实验。其中 CAFS 和 CADU 不需要已经标记的初始训练集;而且采



样数目是 CWMM 和 MEC 的参数,为了保证实验结果的有效性,将采样数目设置为 CAFS、TALK CADU 的计算值。

表 5 显示在 9 个数据集上,CAFS 在大部分数据集上表现

优异,其中平均代价相对于 ALCE,CWMM,MEC,TALK,CADU 算法分别降低了 30%、37%、35%、27%、10%,在平均排名上也取得了最好的成绩。

表 5 CAFS 算法与其他代价敏感主动学习算法在不同数据集上的平均代价对比

Tab. 5 Comparison of average cost of CAFS algorithm and other cost-sensitive active learning algorithms on different datasets

算法	Allmal	Arcene	Banana	Credit6000	Heart	Ionosphere	Madelon	Sonar	Spambase	平均排名
ALCE	0.6944	0.9250	0.4546	0.6486	0.5533	1.4073	1.4634	1.1221	0.9877	3.2222
CWMM	1.9111	0.9250	0.6314	0.6649	0.6659	0.5094	1.4634	1.2990	1.0880	4.4444
MEC	1.5111	0.9250	0.6735	0.6558	0.5695	0.6541	1.4634	1.2702	1.0897	4.2222
TALK	0.6944	1.1200	0.8966	0.3207	0.8889	1.2821	1.0000	0.9327	0.7982	3.8889
CADU	0.8333	0.6850	0.3306	0.4154	0.6630	0.7749	0.9777	0.9087	0.7666	2.4444
CAFS	0.7500	0.7050	0.4226	0.4808	0.6259	0.5128	0.8485	0.6827	0.6881	2.0000

5.5 与非代价敏感学习算法的对比实验

最后,为了实验的完整性,CAFS 与 3 个非代价敏感学习算法——投票熵采样算法 (Vote Entropy Sampling algorithm, VES)、一致熵采样算法 (Consensus Entropy Sampling algorithm, CES) 和最大分歧采样算法 (Max Disagreement Sampling algorithm, MDS) 进行对比。三种算法选取了不同的采样方案,并且有 3 个基本分类器组成,分别是决策树^[26]、随机森林^[27]和带径向基函数 (Radial Basis Function, RBF) 内核

的支持向量机 (Support Vector Machine, SVM)^[28]。因为某些算法在单次实验中会有结果的偏差,所以进行了 5 次实验。实验结果如表 6 所示,由于非代价敏感学习算法不考虑代价因素,所以在为保证实验结果的统一性,计算平均代价时会根据学习结果与代价设置进行代价计算。由表 6 可以看出,CAFS 算法在 4 个算法中平均排名最好,并且平均代价对应于 VES、CES、MDS 算法分别降低了 13.8%、14.34%、19.67%。

表 6 CAFS 算法与其他非代价敏感学习算法在不同数据集上的平均代价的对比

Tab. 6 Comparison of average cost of CAFS algorithm and other cost-insensitive active learning algorithms on different datasets

算法	Allmal	Arcene	Banana	Credit6000	Heart	Ionosphere	Madelon	Sonar	Spambase	平均排名
VES	1.0333	1.0970	0.3999	0.5900	0.6896	0.4148	1.0782	0.8952	0.3947	2.0000
CES	0.7444	0.9990	0.5838	0.6182	0.7193	0.5140	1.1255	0.9452	0.4253	2.8889
MDS	1.1500	1.0370	0.4478	0.6016	0.7844	0.5470	1.1702	0.9183	0.4606	4.0000
CAFS	0.7500	0.7050	0.4226	0.4808	0.6259	0.5128	0.8485	0.6827	0.6881	1.6667

5.6 实验结果分析

综合以上实验结果,有如下结论:

1) CAFS 算法与主流的代价敏感学习 CSLR、CSDT 和 CSRF 相比,平均代价是最低的。

2) CAFS 与同类的代价敏感主动学习算法 CWMM、MEC、TALK 和 CADU 相比,实验结果是最优的。

实验结果表明 CAFS 算法能够有效地降低总代价。

6 结语

本文提出的基于最远总距离采样的主动学习算法,建立了 LUD 模型,并提出了最远总距离采样的策略。利用 3WD 思想使得学习的过程更加完善。标签均匀分布模型在给定的代价以及假设的均匀分布的条件下,可获得最优的采样数目。最远总距离采样策略,综合考虑了信息量和样本的总体特性,使得选择的样本更具代表性。下一步的主要工作包含两个方面:其一是将 LUD 模型推广到多类别的学习任务中;其二是设计更加合适的样本采样策略,进一步减小算法的代价,提高预测精度。

参考文献 (References)

- [1] SETTLES B. Active Learning [M]. San Rafael, CA: Morgan and Claypool Publishers, 2012: 1–114.
- [2] ZHU X, GOLDBERG A B. Introduction to Semi-Supervised Learning [M]. San Rafael, CA: Morgan and Claypool Publishers, 2009: 130.
- [3] SEUNG H S, OPPER M, SOMPOLINSKY H. Query by committee [C]// COLT 1992: Proceedings of the 5th Annual ACM Conference on Computational Learning Theory. New York: ACM, 1992: 287–294.
- [4] COHN D A, CHAHRAMANI Z, JORDAN M I, et al. Active learning with statistical models [J]. Journal of Artificial Intelligence Research, 1996, 4(1): 129–145.
- [5] WANG M, MIN F, ZHANG Z H, et al. Active learning through density clustering [J]. Expert Systems with Applications, 2017, 85: 305–317.
- [6] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2001, 2(1): 45–66.
- [7] THOMPSON C A. Active learning for natural language parsing and information extraction [C]// ICML 1999: Proceeding of the 16th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 1999: 406–414.
- [8] ZHANG C, CHEN T. An active learning framework for content-based information retrieval [J]. IEEE Transactions on Multimedia, 2002, 4(2): 260–268.
- [9] YU D, VARADARAJAN B, DENG L, et al. Active learning and semi-supervised learning for speech recognition: a unified framework using the global entropy reduction maximization criterion [J]. Computer Speech and Language, 2010, 24(3): 433–444.
- [10] MARGINEANTU D D. Active cost-sensitive learning [C]// IJCAI 2005: Proceedings of the 19th International Joint Conference on Ar-



- tificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, 2005: 1622 – 1623.
- [11] MIN F, LIU F L, WEN L Y, et al. Tri-partition cost-sensitive active learning through k NN [J]. Soft Computing, 2017, 23(5): 1557 – 1572.
- [12] WU Y X, MIN X Y, MIN F, et al. Cost-sensitive active learning with a label uniform distribution model [J]. International Journal of Approximate Reasoning, 2019, 105: 49 – 65.
- [13] YAO Y. Three-way decision: an interpretation of rules in rough set theory [C]// Proceedings of the 2009 International Conference on Rough Sets and Knowledge Technology, LNCS 5589. Berlin: Springer, 2009: 642 – 649.
- [14] 李华雄, 周献中, 黄兵, 等. 决策粗糙集与代价敏感分类[J]. 计算机科学与探索, 2013, 7(2): 126 – 135. (LI H X, ZHOU X Z, HUANG B, et al. Decision-theoretic rough set and cost-sensitive classification [J]. Journal of Frontiers of Computer Science and Technology, 2013, 7(2): 126 – 135.)
- [15] 刘盾, 李天瑞, 李华雄. 粗糙集理论: 基于三支决策视角[J]. 南京大学学报(自然科学版), 2013, 49(5): 574 – 581. (LIU D, LI T R, LI H X. Rough set theory: a three-way decisions perspective [J]. Journal of Nanjing University (Natural Science), 2013, 49(5): 574 – 581)
- [16] 杨习贝, 杨静宇. 邻域系统粗糙集模型[J]. 南京理工大学报, 2012, 36(2): 291 – 295. (YANG X B, YANG J Y. Rough set model based on neighborhood system [J]. Journal of Nanjing University of Science and Technology, 2012, 36(2): 291 – 295.)
- [17] SETTLES B, CRAVEN M, Friedland L. Active learning with real annotation costs [EB/OL]. [2018-12-13]. https://www.researchgate.net/publication/228770726_Active_learning_with_real_annotation_costs.
- [18] LIU A, JUN G, GHOSH J. Spatially cost-sensitive active learning [C]// SDM 2009: Proceedings of the 2009 SIAM International Conference on Data Mining. Philadelphia, PA: SIAM, 2009: 814 – 825.
- [19] ZHAO P L, HOI S C H. Cost-sensitive online active learning with application to malicious URL detection [C]// KDD 2013: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 919 – 927.
- [20] CHEN P-L, LIN H-T. Active learning for multiclass cost-sensitive classification using probabilistic models [C]// TAAI 2013: Proceedings of the 2013 Conference on Technologies and Applications of Artificial Intelligence. Washington, DC: IEEE Computer Society, 2013: 13 – 18.
- [21] DEMIR B, MINELLO L, BRUZZONE L. Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method [J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(2): 1272 – 1284.
- [22] HUANG K-H, LIN H-T. A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning [C]// ICDM 2016: Proceedings of the 2016 IEEE 16th International Conference on Data Mining. Piscataway, NJ: IEEE, 2016: 925 – 930.
- [23] BAHNSEN A C, AOUADA D, OTTERSTEN B. Example-dependent cost-sensitive logistic regression for credit scoring [C]// ICMLA 2014: Proceedings of the 2014 13th International Conference on Machine Learning and Application. Washington, DC: IEEE Computer Society, 2014: 263 – 269.
- [24] BAHNSEN A C, AOUADA D, OTTERSTEN B. Example-dependent cost-sensitive decision trees [J]. Expert Systems with Applications, 2015, 42(19): 6609 – 6619.
- [25] BAHNSEN A C, AOUADA D, OTTERSTEN B. Ensemble of example-dependent cost-sensitive decision trees [EB/OL]. [2018-12-13]. <https://arxiv.org/pdf/1505.04637v1.pdf>.
- [26] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81 – 106.
- [27] LIAW A, WIENER M. Classification and regression by random forest [J]. R News, 2002, 2/3: 18 – 22.
- [28] CRISTIANINI N, SHAWE T J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods [M]. Cambridge, Eng.: Cambridge University Press, 2000: 46 – 71.

This work is partially supported by the Scientific Innovation Group for Youths of Sichuan Province (2019JDTD0017), the Applied Basic Research Project of Sichuan Province (2017JY0190).

REN Jie, born in 1996, M. S. candidate. His research interests include active learning.

MIN Fan, born in 1973, Ph. D., professor. His research interests include granular computing, recommender system, active learning.

WANG Min, born in 1980, M. S., associate professor. Her research interests include data mining, active learning.