DOI: 10.11992/tis.201711027
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180411.1021.006.html

SUCE: 基于聚类集成的半监督二分类方法

闵帆, 王宏杰, 刘福伦, 王轩
(西南石油大学 计算机科学学院, 四川 成都 610500)

摘 要: 半监督学习和集成学习是目前机器学习领域中的重要方法。半监督学习利用未标记样本, 而集成学习综合多个弱学习器, 以提高分类精度。针对名词型数据, 本文提出一种融合聚类和集成学习的半监督分类方法 SUCE。在不同的参数设置下, 采用多个聚类算法生成大量的弱学习器; 利用已有的类标签信息, 对弱学习器进行评价和选择; 通过集成弱学习器对测试集进行预分类, 并将置信度高的样本放入训练集; 利用扩展的训练集, 使用 ID3、Nave Bayes、kNN、C4.5、OneR、Logistic 等基础算法对其他样本进行分类。在 UCI 数据集上的实验结果表明, 当训练样本较少时, 本方法能稳定提高多数基础算法的准确性。

关键词: 集成学习; 聚类; 聚类集成; 半监督; 二分类
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673−4785(2018)06−0974−07

中文引用格式: 闵帆, 王宏杰, 刘福伦, 等. SUCE: 基于聚类集成的半监督二分类方法[J]. 智能系统学报, 2018, 13(6): 974–980.
英文引用格式: MIN Fan, WANG Hongjie, LIU Fulun, et al. SUCE: semi-supervised binary classification based on clustering ensemble[J]. CAAI transactions on intelligent systems, 2018, 13(6): 974–980.

SUCE: semi-supervised binary classification based on clustering ensemble

MIN Fan, WANG Hongjie, LIU Fulun, WANG Xuan
(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract: Semi-supervised learning and ensemble learning are important methods in the field of machine learning. Semi-supervised learning utilize unlabeled samples, while ensemble learning combines multiple weak learners to improve classification accuracy. This paper proposes a new method called Semi-sUpervised classification through Clustering and Ensemble learning (SUCE) for symbolic data. Under different parameter settings, a number of weak learners are generated using multiple clustering algorithms. Using existing class label information the weak learners are evaluated and selected. The test sets are pre-classified by weak learners ensemble. The samples with high confidence are moved to the training set, and the other samples are classified through the extended training set by using the basic algorithms such as ID3, Nave Bayes, kNN, C4.5, OneR, Logistic and so on. The experimental on the UCI datasets results show that SUCE can steadily improve the accuracy of most of the basic algorithms when there are fewer training samples.
Keywords: ensemble learning; clustering; clustering ensemble; semi-supervised; binary classification

在机器学习[1]领域中, 半监督学习[2-3]和集成学习[4]是当前的研究热点。它们被广泛应用于智能信息处理[5]、图像处理[6]、生物医学[7]等领域。在许多大数据场景中, 样本属性的获取容易且廉价, 而其标签的获取则困难且昂贵[8]。如果只使用少量已标记样本进行学习, 那么训练得到的分类模型通常会造成过度拟合[9]。为此, Merz 等[10]于 1992 年提出半监督分类, 它不依赖外界交互, 充分利用未标记样本, 有效提高分类模型的稳定性和精度。

集成学习是指先构建多个学习器, 再采用某种集成策略进行结合, 最后综合各个学习器的结果输出最终结果。集成学习中的多个学习器可以

收稿日期: 2017−11−21. **网络出版日期**: 2018−04−11.
基金项目: 国家自然科学基金项目 (61379089).
通信作者: 闵帆. E-mail: minfanphd@163.com.

是同种类型的弱学习器, 也可以是不同类型的弱学习器, 基于这些弱学习器进行集成后获得一个精度更高的“强学习器”[11-12]。

基于聚类的分类算法是指先进行数据聚类[13], 然后根据类簇和标签信息进行分类。其优点是需要的标签较少, 但单一算法的聚类效果不稳定或不符合类标签分布时, 分类效果受到严重影响。2002 年 Strehl 等[14]提出“聚类集成”, 使用不同类型的聚类算法构造不同的学习器, 结合这些学习器可得到更可靠更优的聚类结果; Fred 等[15]提出通过对同一种聚类算法选取不同参数来构造学习器; Zhou[16]利用互信息设定权重, 采用基于投票、加权投票进行聚类集成学习; Zhang[17]提出一种无标签数据增强集成学习的方法 UDEED, 能够同时最大化基分类器在有标签数据上的精度和无标签数据上的多样性。

本文针对名词型数据分类问题, 在半监督学习的框架之下, 融合聚类和集成学习技术, 提出一种新的半监督分类算法 (semi-supervised binary classification based on clustering ensemble, SUCE)。通过在 UCI 4 个数据集上的实验表明, 该方法比传统的 ID3、kNN、C4.5 等算法的分类效果要好。而且, 当标签较少时, 其分类优势更为明显。

1　基本概念

分类问题的基础数据为决策系统。

定义 1[18]　决策系统 S 为一个三元组:

$$S=(U,C,d) \tag{1}$$

式中: U 是对象集合也称为论域; C 是条件属性集合; d 是决策属性。本文只研究名词型数据的二分类问题, 所以决策属性只有两个属性值即 $|V_d|=2$。一般假设所有的条件属性值已知, 而仅有部分样本决策属性值已知。这些对象构成了训练集 U_r, 而 $U_t=U-U_r$ 构成了测试集。实际上, 在半监督学习中, 测试集的对象也参与了训练模型的构建。

聚类问题不涉及决策属性 d。聚类集成是指关于一个对象集合的多个划分组合成为一个统一聚类结果的方法, 目标就是要寻找一个聚类, 使其对于所有的输入聚类结果来说, 尽可能多地符合[19]。

如图 1 所示, 聚类集成的过程为: 首先对 $U=\{x_1,x_2,\cdots,x_m\}$, 通过集成学习器集合 $\mathbb{C}=\{C_1,C_2,\cdots,C_k\}$, 得到 $P=\{p_1,p_2,\cdots,p_k\}$, 其中 $p_i(i=1,2,\cdots,k)$ 为第 i 个聚类学习器得到的聚类结果。最后通过一致性函数对 $x\in U_t$ 的 k 个预测标签进行集成, 得到一个统一的集成标签 $h(x)$。

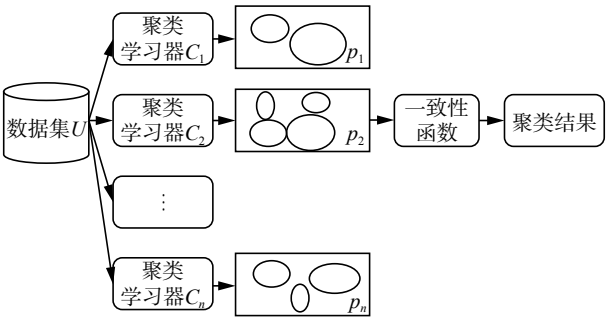

图 1　聚类集成过程示意图

Fig. 1　The diagram of clustering ensemble

集成学习中, 学习器之间的差异性被认为是影响集成结果的关键因素之一[20]。聚类集成的第一步是通过不同类型聚类基学习器产生多个聚类结果, 从不同的方面反映数据集的结构, 有利于集成[21]。在本文中, k-Means[22]、EM[23]、FarthestFirst[24]和 HierarchicalClusterer[25]4 个聚类算法将作为聚类集成的基础学习算法, 并且每次运行都设置不同的参数。k-Means 原理简单运行速度较快, 但依赖于初始参数设置使得聚类结果存在不稳定性, 并且不能有效针对非凸形状分布数据聚类。EM 不需要事先设定类别数目, 计算结果稳定、准确, 但算法相对复杂, 收敛较慢不适用于大规模数据集和高维数据。HierarchicalClusterer 没有任何目标函数, 簇合并后不可逆转, 将局部最优作为全局最优解, 聚类结果依赖于主观获得。FarthestFirst 在迭代过程中减少待聚类样本数和类别数, 具有精简聚类结果的效果。每个算法各有优劣, 适用的场景不同; 因此需要对它们进行集成化来实现优势互补。因为本文只研究名词型数据的二分类问题, 所以在聚类时, 聚簇的数量直接设为类别数量, 在实验中, 本文将所有聚类算法的聚簇数量设定为 2。

聚类效果的主要评价指标有 JC 系数、FM 指数、DB 指数和 DI 指数等。本文通过聚类方法研究二分类问题, 使用 U_r 的聚类纯度对聚类结果进行评估。通常来说, 聚类纯度越高则表明聚类效果越好。

定义 2　聚类纯度 (purity of cluster, PC)

设数据集 $U=U_t\cup U_r$, 对于任意聚类学习器 $C\in\mathbb{C}$ 的聚类结果 $\boldsymbol{t}(U)=[\boldsymbol{t}(U_t)\ \boldsymbol{t}(U_r)]$, 其中 $\boldsymbol{t}(U_r)=[t(x_1)\ t(x_2)\ \cdots\ t(x_{|U_r|})]$, 用 $\boldsymbol{d}(U_r)=[d(x_1)\ d(x_2)\ \cdots\ d(x_{|U_r|})]$ 表示 $x_i\in U_r$ 的真实标签。

那么基学习器 C 对于 U_r 的聚类纯度可表示为

$$\mathrm{PC}(U_r)=\frac{1}{|U_r|}\sum_{i=1}^{|U_r|}\sigma(t(x_i),d(x_i)) \tag{2}$$

式中：

$$\sigma(a,b)=\begin{cases}1, & a=b\\ 0, & \text{其他}\end{cases} \tag{3}$$

另外，聚类集成学习存在一个必须要解决的问题：簇标签与真实标签的对应。

定义 3　(标签对应) 任意聚类基学习器 $C\in\mathbb{C}$，根据对训练集 U_r 上的聚类纯度 $\mathrm{PC}(U_r)$，得到 $x\in U_t$ 中样本的聚类标签。标签对应函数 $A(U_t)$ 可定义为

$$A(U_t)=\begin{cases}\textbf{normal}(U_t), & \mathrm{PC}(U_r)>\theta\\ \textbf{covert}(U_t), & \mathrm{PC}(U_r)<1-\theta\\ \mathrm{reset}(U_t), & 1-\theta<\mathrm{PC}(U_r)<\theta\end{cases} \tag{4}$$

式中：

$$\textbf{normal}(U_t)=[\sigma(d'(x_1)\ t(x_1))\ \cdots\ \sigma(d'(x_{|U_t|})\ t(x_{|U_t|}))],$$
$$\textbf{convert}(U_t)=[\sigma(d'(x_1)\ 1-t(x_1))\ \cdots\ \sigma(d'(x_{|U_t|})\ 1-t(x_{|U_t|}))],$$
$$\mathrm{reset}(U_t)=-1_{U_t\times 1},\text{且}x_i\in U_t,i=1,2,\cdots,|U_t|。$$

本文用 $t(x)$ 和 $d'(x)$ 分别表示样本 $x\in U_t$ 的聚类标签和预测标签。θ 是用户设置的阈值，当 $\mathrm{PC}(U_r)>\theta$ 时，即表示聚类标签与类标签相匹配，将调用 $\mathrm{normal}(U_t)$ 函数，并直接把聚类标签作为预测标签；当 $\mathrm{PC}(U_r)>\theta$ 时，即表示聚类标签与类标签相反，将调用 $\mathrm{covert}(U_t)$ 函数，把聚类标签取反后作为预测标签；当 $\mathrm{PC}(U_r)$ 介于 $1-\theta$ 和 θ 之间，即认为聚类结果不适于指导标签预测，调用 $\mathrm{reset}(U_t)$ 函数，用 -1 表示 $x\in U_t$ 的预测标签。

例 1　对 $U_r=\{x_1,x_2,x_3\}$，$U_t=\{y_1,y_2\}$，有 $U=U_r\cup U_t$，且 $\boldsymbol{d}(U_r)=[1\ 0\ 1]$，$\theta$=0.9。

1) 如果 $\boldsymbol{t}(U)=[1\ 0\ 1\ 10]$，$\mathrm{PC}(U_r)=1>\theta$，所以 $\boldsymbol{d}'(U_t)=\textbf{normal}(U_t)=[1\ 0]$；

2) 如果 $\boldsymbol{t}(U)=[0\ 1\ 0\ 0\ 1]$，$\mathrm{PC}(U_r)=0<1-\theta$，所以 $\boldsymbol{d}'(U_t)=\textbf{convert}(U_t)=[1\ 0]$；

3) 如果 $\boldsymbol{t}(U)=[0\ 0\ 0\ 0\ 1]$，$1-\theta<\mathrm{PC}(U_r)=0<1-\theta$，所以 $\boldsymbol{d}'(U_t)=\mathrm{reset}(U_t)=[-1\ -1]$。

聚类学习器集合 $\mathbb{C}$ 将给样本 $x\in U_t$ 标记 $|\mathbb{C}|$ 个聚类标签，并根据定义 (2) 和定义 (3) 得到 $|\mathbb{C}|$ 个预测标签。

定义 4　(一致性) 聚类学习器 $C_i\in\mathbb{C}$ 对 $x\in U_t$ 上的预测标签 $d_i'(x)$，且 $d_i'(x)\in\{0,1\}$，那么集成标签 $h(x)$ 的值为

$$h(x)=\begin{cases}d'(x), & \sum_{i=1}^{|\mathbb{C}|}d_i'(x)=|C|\ \text{or}\ \sum_{i=1}^{|\mathbb{C}|}d_i'(x)=0\\ -1, & \text{其他}\end{cases} \tag{5}$$

例 2　采用与例 1 中相同的 U_t 和 U_r，且 $|\mathbb{C}|$=3，若 C_1 的预测标签 $\boldsymbol{d}_1'(U_t)=[1\ 0]$，若 C_2 的预测标签 $\boldsymbol{d}_2'(U_t)=[1\ 1]$，若 C_3 的预测 $\boldsymbol{d}_3'(U_t)=[1\ 0]$。

对 $U_t=\{y_1,y_2\}$ 的结果为

因为 $\sum_{i=1}^{3}d_i'(y_1)=|\mathbb{C}|=3$，所以 $h(y_1)$=1；

因为 $\sum_{i=1}^{3}d_i'(y_2)=2$，所以 $h(y_2)$=−1。

2　算法设计与分析

本节首先描述算法的总体框架，然后进行算法伪代码描述，最后分析算法复杂度。

2.1　算法总体方案

基于集成的半监督分类方法主要是通过集成学习控制无标记样本的标注过程来减少未标记的不确定性[12]。然而，目前在利用集成学习辅助半监督学习方面的方法研究较少，主要是存在如下矛盾：半监督学习适用于标记样本不足的情况，然而传统的集成学习本身就需要大量的标记样本进行训练[12]。针对上述问题，SUCE 综合聚类集成与半监督学习，在已知标签较少的情况下，有效提高分类器的精度。

如图 2 所示，基于聚类集成的半监督分类过程为：第 1 个分图说明，首先通过聚类集成，将 B 中部分没有类别样本 C 的类标签预测出来；达到“扩大”有类别的样本集合 (A 变成了 A+C)，“缩小”了未标记类别集合 (B 变成了 B')。第 2 个分图说明，对于扩大后的集合 (A+C) 利用分类模型，完成预测没有类别的样本 B'。

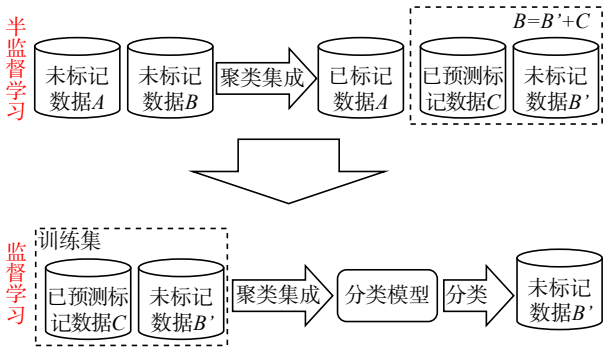

图 2　基于聚类集成的半监督分类示意图

Fig. 2　The diagram of semi-supervised classification based on clustering ensemble

2.2　算法描述

在训练阶段，本算法将依次对数据集进行 4 步处理，从而生成分类器：

1) 通过 $\mathrm{getLabel}(U_r)$ 获取训练集 U_r 的标签 $\boldsymbol{L}_{U_r}$。然后，利用 $\mathrm{remove}(U_r)$ 对 U_r 去标签得到 U_r'；并将 $U_r'\cup U_t$ 得到无标签 U。

2) 通过多个基于 EM、*K*-Means、Farthest-First 和 HierarchicalClusterer 等聚类算法的个体学习器对 U 进行全局聚类。根据已获取的 $\boldsymbol{L}_{U_r}$, 计算第 i 个聚类学习器 $\boldsymbol{L}_l$ 在 U_r 上的聚类纯度 $PC(i)$。如果 $PC(i)$ 高于阈值 θ, $\boldsymbol{L}_l$ 将继续参加集成学习, 并将 $\boldsymbol{L}_l$ 移入到学习器集合 E 中即 $E \cup \boldsymbol{L}_l$。

3) 对测试集的预测标签进行集成学习。通过 $h(x)$ 一致性函数依次对测试集每个样本 $x \in U_t$ 的预测标签进行一致性处理。如果 E 中所有学习器对 x 的预测标签均一致, 将预测标签 $d'(x)$ 赋给 x 得到 $x'=(x, d'(x))$。x' 移入到训练集 $U_r \cup \{x'\}$, 同时在测试集中将其删除 $U_t-\{x\}$。

4) 对扩大规模后的 U_r 进行学习, 再对缩减规模后的 U_t 进行分类 $\boldsymbol{L}_{U_t}$=classifier(U_r, U_t) 得到 U_t 的类标签 $\boldsymbol{L}_{U_t}$; 然后, 获取 U_r 的标签 $\boldsymbol{L}_{U_r}$=getLabel(U_r)。最终得到 U 类标签 $\boldsymbol{L}_U$=combine($\boldsymbol{L}_{U_t}$, $\boldsymbol{L}_{U_r}$)。

SUCE: 基于集成聚类的半监督分类算法

算法　**SUCE**

输入　训练集 U_r, 测试集 U_t, 阈值;

输出　U_t 的类标签向量 $\boldsymbol{L}_{U_t}$。

优化目标: 最大化分类精度;

1) $U=\varnothing$, $E=\varnothing$; //初始化
2) $\boldsymbol{L}_{U_r}$=getLabel(U_r) //获取 U_r 类标签
3) $U_r' = \text{remove}(U_r)$; //隐藏 U_r 类标签
4) $U = U_r' \cup U_t$;
5) $\boldsymbol{L}_1 = \text{KMeans}(U)$;
6) $\boldsymbol{L}_2 = \text{EM}(U)$;
7) $\boldsymbol{L}_3 = \text{FarthestFirst}(U)$;
8) $\boldsymbol{L}_4 = \text{HierarchicalCluster}(U)$;
9) for (i=0; i<4; i++) do //筛选基学习器
10)　if ($PC(i)>\theta$) then
11) $E \cup \boldsymbol{L}_l$
12) end if
13) end for
14) for (each $x \in U_t$) do //标签一致性处理
15) if ($h(x) = d'(x)$) then
16) $\boldsymbol{L}_{U_t(x)} = d'(x)$;
17) else then
18) $\boldsymbol{L}_{U_t(x)} = -1$;
19) end if
20) end for
21) for (each $x \in U_t$) do //扩充训练集
22) if ($\boldsymbol{L}_{U_t(x)}$! = −1) then
23) $x'=(x, d'(x))$
24) $U_r \cup \{x'\}$;
25) $U_t-\{x\}$;
26) end if
27) end for
28) $\boldsymbol{L}_{U_t}$=classifier(U_r, U_t); //分类
29) $\boldsymbol{L}_{U_r}$=getLabel(U_r);
30) $\boldsymbol{L}_U$=combine($\boldsymbol{L}_{U_t}$, $\boldsymbol{L}_{U_r}$);

2.3　复杂度分析

为方便讨论, 假设训练集 U_r 的对象数量为 n, 条件属性数量为 c, 测试集 U_t 的对象数量为 m。基学习器数量为 $|E|$, 迭代次数为 t、聚类簇数为 k。SUCE 算法细分为以下 4 个阶段。

1) 对数据集进行去标签化预处理。在隐藏 U_r 类标签之前, 需先记录其真实类标签, 如第 2) 行所示再隐藏 U_r 中的类标签, 如第 (3) 行所示。至此, 需要对 U_r 进行两次遍历, 共执行 $2n$ 次计算。接下来是合并去标签后的 U_r 和 U_t, 构建无标签论域 U。第 1 阶段, 计算机将共执行 $3n+m$ 次运算, 故该阶段的时间复杂度为 $O(n+m)$。

2) 分别通过基于 *K*-Means、EM、Farthest-First 和 HierarchicalClusterer 基学习器对 U 进行全局聚类, 如第 5)~8) 行所示。其时间复杂度分别为 $O(kt(n+m))$、$O(ct(n+m))$、$O(k(n+m))$、$O((n+m)^2 \lg(n+m))$, 然后计算基学习器的聚类纯度, 并对其进行筛选, 共执行 $n \times |E|$ 次运算, 如第 9)~13) 行所示。所以, 第 2 阶段的时间复杂度为 $O((n+m)(ct+(n+m)\lg(n+m)))$。

3) 对 U_t 中的对象进行一致化处理。遍历 U_t 中对象, 共执行 m 次处理, 如第 14)-20) 行所示。然后将 U_t 中置信度高的对象移入到 U_r, 如第 21)~27) 行所示, 共执行 $2m$ 次计算, 故时间复杂度为 $O(m)$。

4) 对扩展后的 U_r 进行学习, 并对 U_t 进行分类。该阶段的时间复杂度根据所采用的具体分类算法变化而变化。

综上, 因为第 1、3、4 阶段的时间复杂度远小于第 2 阶段。所以, SUCE 的时间复杂度为 $O((n+m)(ct+(n+m)\lg(n+m)))$。为方便表述, 数据集规模 $n+m$ 改写成 $|U|$, SUCE 的时间复杂度为 $O(|U|(ct+|U|\lg(|U|)))$。

3　实验及分析

本节通过实验回答以下 3 个问题: 1) 如何设置合适的 θ 阈值; 2) SUCE 应用于哪些基础算法效果更好; 3) 相比于流行的分类算法, SUCE 能否提高分类器的精度。

3.1　实验设置

实验采用了 UCI 数据库中的 Sonar、Iono-

sphere、Wdbc 和 Voting4 个数据集。Sonar、Wdbc 是连续型数据，因此通过 Weka 应用默认方法对其进行离散化处理。

根据 UCI 数据集的样本数量，实验设置的训练集规模分别为 2%、4%、6%、8%、10%、12%、14% 和 16%。在测试集中，样本标签不可见，直到所有的未分类样本都得到预测标签。为减小实验随机误差，每个结果均为 200 次相同实验的平均值。所有 (对比) 实验均采用上述相同的实验参数，如表 1 所示。

表 1　数据集的描述

Table 1　Description of the data set

序号	数据集	样本数	特征数	类别数
1	Sonar	208	61	2
2	Wdbc	569	31	2
3	Ionosphere	351	35	2
4	Voting	435	17	2

3.2　实验结果与分析

图 3 显示了 Sonar、Wdbc、Ionosphere 和 Voting 数据集在不同阈值 θ 和训练集规模下的平均分类精度变化。通过实验数据观察发现，θ=0.8 左右时，SUCE 在 4 个数据集上均能取得最好的分类效果。在 Sonar 和 Voting 数据集上，对于不同的 θ 取值，随着训练集规模的扩大，平均分类精度会呈现出先增加后趋于稳定的趋势。因为随着阈值 θ 的提高，筛选过后还保留的个体学习器通常会变得更少，所以获得的样本标签并没有提高，从而导致分类效果没有提升。对于 Ionosphere 和 Wdbc，训练集规模并不太影响平均分类精度。

表 2 显示了 SUCE 作用在 ID3、J48、Bayes、*k*NN、Logistic、OneR 等基础算法上，并对 Sonar、Wdbc、Ionosphere 和 Voting 数据集进行半监督分类的分类结果。实验参数设置为：θ=0.8，训练集比例=4%。Win 值的计算如下：在某一数据集上，如果某种算法效果比其对比算法精度高 1% 以上，则该算法得 1 分；否则两种算法效果相当且均不得分。

通过表 2 可以统计发现，SUCE 获胜 14 次，打平 5 次，失败 5 次。在 Sonar、Wdbc 和 Ionosphere 数据集上的分类效果要优于基础算法。但 SUCE 在 Voting 数据集上对基础算法分类效果的提升不明显。

SUCE 更适用于 ID3、C4.5、OneR 等基础算法。例如，在 Sonar 数据上，SUCE-C4.5 获得了高达 14% 的精度提升。然而，SUCE 对 Naive Bayes 算法的改进不明显。

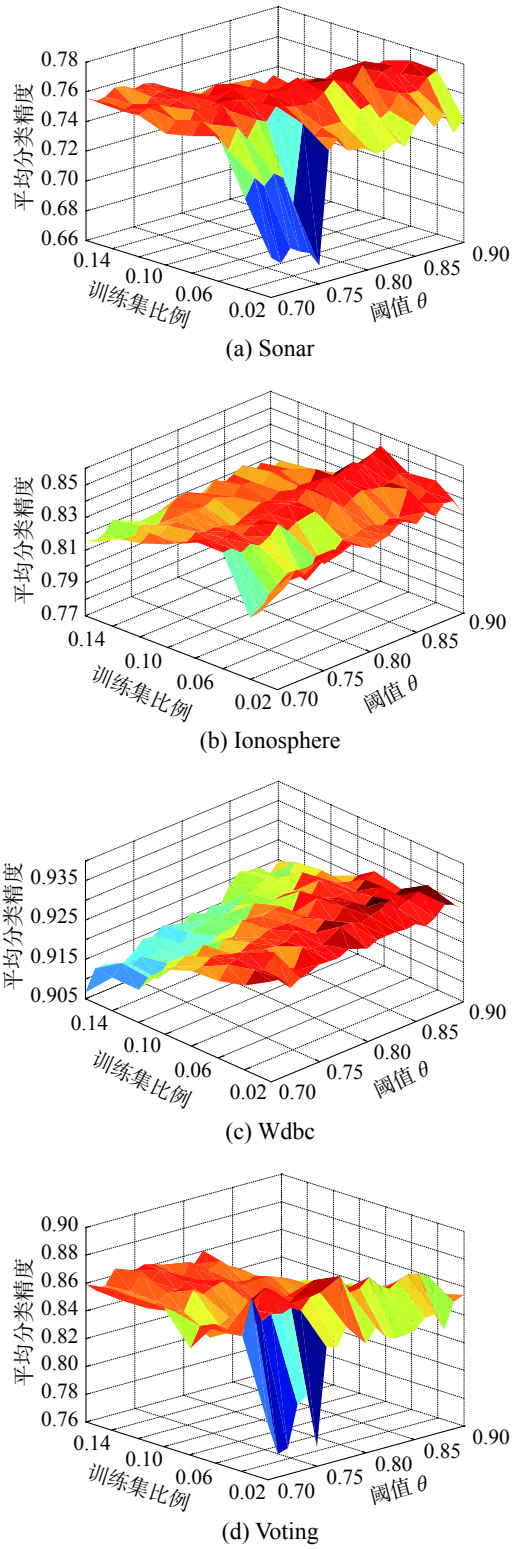

图 3　SUCE-ID3 在不同数据集上的分类比较

Fig. 3　The diagram of comparison of SUCE-ID3 classification on different datasets

现在可以回答本节提出的问题。1) 取为 0.8 左右较合适；2) SUCE 应用于 ID3、C4.5、OneR 等基础算法效果更好；3) 相比基础算法，SUCE 通常可以提高分类器的精度。

表 2　SUCE 与基础算法分类精度对比

Table 2　Comparing the classification accuracy of SUCE and basic algorithms

算法	数据库	Sonar	Wdbc	Ionosphere	Voting	Win
ID3	Initial	0.634 75±0.014 27	0.878 89±0.001 73	0.748 81±0.006 82	0.878 59±0.005 06	0
	SUCE-ID3	0.759 38±0.006 60	0.926 42±0.000 02	0.821 84±0.002 64	0.871 36±0.003 01	3
C4.5	Initial	0.612 56±0.018 41	0.863 16±0014 00	0.750 89±0.011 18	0.904 19±0.006 78	1
	SUCE-C4.5	0.752 13±0.006 91	0.915 45±0.000 01	0.819 90±0.002 32	0.869 92±0.003 58	3
Naive Bayes	Initial	0.771 82±0.005 96	0.949 54±0.000 03	0.884 13±0.002 26	0.893 18±0.001 38	1
	SUCE-Bayes	0.783 88±0.005 93	0.942 69±0.000 02	0.875 40±0.001 82	0.865 02±0.000 74	1
kNN	Initial	0.686 25±0.010 37	0.929 89±0.000 05	0.824 18±0.002 09	0.903 71±0.001 97	1
	SUCE-*k*NN	0.784 38±0.007 66	0.935 65±0.000 04	0.861 75±0.002 03	0.857 01±0.000 88	2
Logistic	Initial	0.666 52±0.009 33	0.913 16±0.000 02	0.849 11±0.001 83	0.891 15±0.002 07	1
	SUCE-Logistic	0.779 88±0.007 03	0.929 16±0.000 02	0.841 13±0.001 81	0.848 03±0.001 27	2
OneR	Initial	0.627 55±0.014 24	0.877 15±0.001 24	0.758 90±0.005 50	0.905 86±0.006 47	1
	SUCE-OneR	0.762 13±0.006 12	0.908 96±0.000 04	0.826 74±0.002 05	0.872 32±0.004 03	3

4　结束语

本文提出的基于集成聚类的半监督二分类算法 SUCE 解决了样本过少情况下的分类效果较差的问题。优点在于通过集成聚类的学习充分挖掘大量未标记样本中的重要信息, 而不需要去求助外界来解决, 降低了学习的成本。在未来的工作中, 进一步研究以下 3 个方向: 1) 由目前只能解决二分类问题过渡到多分类问题; 2) 加入更多学习能力强的聚类算法, 扩大集成学习个体学习器的规模; 3) 引入代价敏感, 增强集成学习的能力。

参考文献:

[1] MITCHELL T M. 机器学习[M]. 曾华军, 张银奎, 译. 北京: 机械工业出版社, 2003.

[2] ZHU Xiaojin. Semi-supervised learning literature survey[R]. Madison: University of Wisconsin, 2008: 63–77.

[3] 张晨光, 张燕. 半监督学习[M]. 北京: 中国农业科学技术出版社, 2013.

[4] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[5] NIGAM K, MCCALLUM A K, THRUN S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine learning, 2000, 39(2/3): 103–134.

[6] SONG Yangqiu, ZHANG Changshui, LEE J, et al. Semi-supervised discriminative classification with application to tumorous tissues segmentation of MR brain images[J]. Pattern analysis and applications, 2009, 12(2): 99–115.

[7] FENG Wei, XIE Lei, Zeng Jia, et al. Audio-visual human recognition using semi-supervised spectral learning and hidden Markov models[J]. Journal of visual languages and computing, 2009, 20(3): 188–195.

[8] SHAHSHAHANI B M, LANDGREBE D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE transactions on geoscience and remote sensing, 1994, 32(5): 1087–1095.

[9] 梁吉业, 高嘉伟, 常瑜. 半监督学习研究进展[J]. 山西大学学报: 自然科学版, 2009, 32(4): 528–534.
LIANG Jiye, GAO Jiawei, CHANG Yu. The research and advances on semi-supervised learning[J]. Journal of Shanxi university: natural science edition, 2009, 32(4): 528–534.

[10] MERZ C J, ST CLAIR D C, BOND W E. Semi-supervised adaptive resonance theory (SMART2)[C]//Proceedings of 1992 International Joint Conference on Neural Networks. Baltimore, USA, 1992: 851–856.

[11] VEGA-PONS S, RUIZ-SHULCLOPER J. A survey of clustering ensemble algorithms[J]. International journal of pattern recognition and artificial intelligence, 2011, 25(3): 337–372.

[12] 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述[J]. 计算机科学, 2017, 44(6A): 7–13.
CAI Yi, ZHU Xiufang, SUN Zhangli, et al. Semi-supervised and ensemble learning: a review[J]. Computer science, 2017, 44(6A): 7–13.

[13] 曾令伟, 伍振兴, 杜文才. 基于改进自监督学习群体智能 (ISLCI) 的高性能聚类算法[J]. 重庆邮电大学学报: 自然科学版, 2016, 28(1): 131–137.
ZENG Lingwei, WU Zhenxing, DU Wencai. Improved Self supervised learning collection intelligence based high

performance data clustering approach[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 28(1): 131–137.

[14] STREHL A, GHOSH J. Cluster ensembles—a knowledge reuse framework for combining partitionings[J]. Journal of machine learning research, 2002, 3: 583–617.

[15] FRED A L N, JAIN A K. Data clustering using evidence accumulation[C]//Proceedings of the 16th International Conference on Pattern Recognition. Quebec, Canada, 2002: 276–280.

[16] ZHOU Zhihua. Ensemble Methods: Foundations and Algorithms[M]. Boca Raton: Taylor and Francis Group, 2012: 135–156.

[17] ZHANG Minling, ZHOU Zhihua. Exploiting unlabeled data to enhance ensemble diversity[J]. Data mining and knowledge discovery, 2013, 26(1): 98–129.

[18] MIN Fan, HU Qinghua, ZHU W. Feature selection with test cost constraint[J]. International journal of approximate reasoning, 2014, 55(1): 167–179.

[19] GIONIS A, MANNILA H, TSAPARAS P. Clustering aggregation[M]//SAMMUT C, WEBB G I. Encyclopedia of Machine Learning. Boston: Springer, 2011.

[20] 罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报, 2007, 30(8): 1315–1324.
LUO Huilan, KONG Fansheng, LI Yixiao. An analysis of diversity measures in clustering ensembles[J]. Chinese journal of computers, 2007, 30(8): 1315–1324.

[21] 杨草原, 刘大有, 杨博, 等. 聚类集成方法研究[J]. 计算机科学, 2011, 38(2): 166–170.
YANG Caoyuan, LIU Dayou, YANG Bo, et al. Research on cluster aggregation approaches[J]. Computer science, 2011, 38(2): 166–170.

[22] 杨玉梅. 基于信息熵改进的 K-means 动态聚类算法[J]. 重庆邮电大学学报: 自然科学版, 2016, 28(2): 254–259.
YANG Yumei. Improved K-means dynamic clustering algorithm based on information entropy[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 28(2): 254–259.

[23] JAMSHIDIAN M, JENNRICH R I. Standard errors for EM estimation[J]. Journal of the royal statistical society. series B, 2000, 62(2): 257–270.

[24] DEEPSHREE A V, YOGISH H K. Farthest first clustering in links reorganization[J]. International journal of web and semantic technology, 2014, 5(3): 17–24.

[25] RASHEDI E, MIRZAEI A. A hierarchical clusterer ensemble method based on boosting theory[J]. Knowledge-based systems, 2013, 45: 83–93.

作者简介：

闵帆，男，1973 年生，教授，博士生导师，主要研究方向为粒计算、代价敏感学习、推荐系统，主持国家自然科学基金 1 项。发表学术论文 100 余篇，被 SCI 检索 30 余篇。

王宏杰，男，1992 年生，硕士研究生，主要研究方向为粒计算、代价敏感学习。发表学术论文 7 篇，其中被 EI 检索 1 篇。

刘福伦，男，1993 年生，硕士研究生，主要研究方向为代价敏感学习、粗糙集。发表学术论文 5 篇，其中被 SCI 检索 2 篇，被 EI 检索 1 篇。