一种适用于不同分类器的样本约简算法

程汝峰，梁永全，刘　彤
（山东科技大学 计算机科学与工程学院，山东 青岛 266590）

摘　要：现有的样本约简算法多数是针对某种分类器设计的，在实际应用中有一定的局限性。结合聚类算法的思想，设计了一种适用于不同分类器的样本约简算法，核心是选取密度高且距离相对较远的样本点。与其他样本约简算法相比较，该算法可以根据需求获得任意大小的样本子集，并适用于多种分类算法；而对包含噪声点的样本集，算法的分类精度和稳定性均有一定程度的提高。

关键词：样本约简；准则；密度；样本子集

中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1672-3767(2017)03-0114-09

An Instance Reduction Algorithm for Different Classifiers

CHENG Rufeng，LIANG Yongquan，LIU Tong
(College of Computer Science and Engineering，Shandong University of Science and Technology，Qingdao，Shandong 266590，China)

Abstract：Most of the existing instance reduction algorithms are designed for particular classifiers，which have some limitations in practical application. Combined with the idea of clustering algorithm，this paper proposes an instance reduction algorithm applicable to different classifiers. The core idea is to select instances with high density and relatively far distance. Compared with other instance reduction algorithms，this algorithm can obtain the instance subset in any size and can be applied to a variety of classification algorithms. For the set with noise，both the classification accuracy and stability of the proposed algorithms can be improved to some extent.

Key words：instance reduction；criterion；density；instance subset

一般来说，不同样本的重要程度是不同的。一些冗余和噪音数据不仅造成大量的存储耗费，而且还会影响学习精度。因此更倾向于根据一定的性能标准，选择代表性样本形成原样本空间的一个子集，之后在这个子集上进行学习。在保持某些性能的基础上，最大限度地降低时间、空间的耗费。

样本约简算法根据性能的要求大致可分为增强型、保持型和混合类型三类。增强型算法的代表有 ENN (edited nearest neighbor)[1]、RENN(repeated ENN)[2]、AKNN(aggregate k nearest neighbor)[3]等；保持型算法的代表有 CNN(condensed nearest neighbor)[4]、RNN(reduced nearest neighbor)[5]、MCS(minimal consistent set)[6]、FCNN(fast nearest neighbor condensation)[7]等；混合型算法的代表性工作有 ICF(iterative case filtering algorithm)[8]、DROP3(decremental reduction optimization procedure)[9]等。文献[10]和文献[11]对样本约简算法进行了很好的综述。

收稿日期：2016-10-12
基金项目：山东省高等学校科技计划项目(J14LN33)；中国博士后科学基金项目(2014M561949)；2014 青岛市博士后项目；山东科技大学研究生科技创新项目(SDKDYC170340)
作者简介：程汝峰(1992—)，男，山东德州，硕士研究生，主要从事数据挖掘与机器学习研究.
梁永全(1968—)，男，山东聊城，教授，博士生导师，主要从事分布式人工智能、数据挖掘与机器等的学习研究，本文通信作者. E-mail：lyq@sdust. edu. cn

近几年，研究者尝试用不同的方法来实现样本约简。针对支持向量机，Chen 等[12]提出一种可以加速支持向量机训练的样本约简算法；针对贝叶斯分类器，Pabitra 等[13]提出一种多尺度的样本约简算法；基于保留区分数据超平面和最大化间隔的思想，文献[14]提出了一种间隔最大化的样本约简算法；基于超矩形聚类算法的思想，文献[15]提出了一种能够有效剔除非边界数据、保留边界和邻近边界数据的样本约简算法；基于谱图理论，文献[16]提出了一种能够有效划分数据集边界和内部实例的谱图样本约简算法；基于 K-MEANS 聚类算法的思想，文献[17]提出了一种能够处理较大数据集的高效样本约简算法；基于部分剔除的思想，文献[18]提出了一种仅剔除可疑噪声数据部分属性的部分样本约简算法。

通过分析可以发现，现有的样本约简算法或者是针对某种分类器设计，或者是倾向于选择边界数据。由于边界数据通常较稀疏并且很可能存在噪声，对原始样本集的代表性较差，很难适用于不同的分类算法。受文献[19]的启发，结合已有的约简准则，本文提出了一种新的样本点选取标准，该标准不针对于某种特定分类器，核心思想是选取密度高且相互之间距离相对较远的样本点。实验表明该标准在多种分类器上都有好的表现，有效地提高了样本约简效果。

1 理论基础

1.1 已有的约简准则

已有的约简准则多数为不确定性准则，通常是选择不确定性信息量大的样本。不确定性的度量方法有多种，相对应的准则也有多个。

最小置信度准则[20]是用概率学习模型计算或估计样本的后验概率。最大熵准则[21]是用信息熵度量样例的不确定性。投票熵准则[22]是综合若干个概率学习模型度量不确定性。一致性约简准则的核心概念是最近异类近邻子集[6]，最小一致子集样本约简算法(MCS)就是用这种度量标准来选择样例，生成最小一致子集。

1.2 精英点约简准则

已有的约简准则忽视了样本的分布信息。本文基于密度和距离的方法提出新的样本点选取准则。在这里称原始样本集中的数据为原数据，被选择出来的样本子集中的数据为精英点。以支持向量机(support vector machine，SVM)[23]为例，如果选择了具有代表性的数据，则对应的分类超平面应该非常接近“正中间”，并且能够有效排除噪声点的影响。从图 1 可以看出，通过精英点训练得出的分类器更具有通用性。

图 2[14]表示的是一组按密度排序的样本分布图，图中数字是按密度降序后各个样本的编号。从图 2 中可以看出，如果仅考虑密度因素，样本点 1～9 都比样本点 10 的密度大，故选取的前 9 个精英点为样本点 1～9。显然这样的选取过于集中，对整个样本集的代表性较差，如果存在重复的数据则效果更差。为了使选择的精英点更具有代表性，则需要更加分散，因此需要考虑距离因素。根据以上分析，易知精英点应同时具有两个特点：精英点本身的密度要高；精英点与其他密度较大的数据点之间的距离要大。

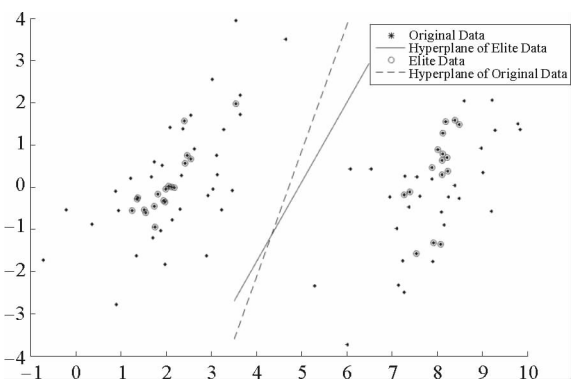

图 1 精英点和相对应的分超平面

Fig. 1 Elite data points and corresponding classified hyperplane

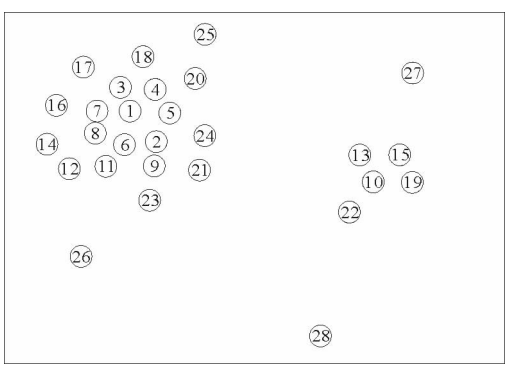

图 2 一组按密度排序的样本分布图

Fig. 2 A set of instance distribution according to the density sort

1.2.1 密度的计算

给定样本集 $D=\{x_i\}_{i=1}^{n}$，$I_D=\{1,2,\cdots,n\}$ 为相应的指标集，$d_{ij}=dist(x_i,x_j)$ 表示数据点 x_i 和 x_j 之间的距离。对于 D 中的任何数据点 x_i，定义局部密度 ρ_i 和高密度点距离 δ_i 两个量。

局部密度 ρ_i 包括截断核和高斯核两种计算方式：

截断核：$\rho_i=\sum_{j\in I_D\backslash\{i\}}\chi(d_{ij}-d_c)$，高斯核：$\rho_i=\sum_{j\in I_D\backslash\{i\}}e^{-(\frac{d_{ij}}{d_c^2})}$。

其中函数：$\chi(x)=\begin{cases}1, & x<0;\\ 0, & x\geqslant 0,\end{cases}$；$I_D\backslash\{i\}$ 表示从指标集 I_D 中剔除 i 指标后得到的集合；参数 $d_c>0$ 为截断距离，其值可以直接给出，也可以通过数据之间的距离来获得；有序距离集合 $D_d=\{d_{(1)},d_{(2)},\cdots,d_{(n)}\}$，该集合通过对所有的 d_{ij} 升序排序得到，其中 $d_{(i)}$ 表示位于第 i 个位置的距离值。给出截取位置所占总体比例 P_{dc}，通过公式 $d_c=d_{\lceil n\times p_{dc}\rceil}$ 能够计算得到 d_c 的值；局部密度 ρ_i 表示 D 中与 x_i 之间的距离小于 d_c 的数据点的个数。

截断核为离散型而高斯核为连续型，因此相对来说后者产生冲突的概率更小。此外对于高斯核，如果与 x_i 的距离小于 d_c 的数据点越多，则 ρ_i 的值越大。

1.2.2 高密度点距离

精英点与其他密度较大的数据点之间的距离要大的特点，要求重新定义高密度点的距离。指标集 $I_D^i=\{k\in I_D:\rho_k>\rho_i\}$，则高密度点距离 δ_i 的定义为：

$$\delta_i=\begin{cases}\min\limits_{j\in I_D^i}\{d_{ij}\}, & I_D^i\neq\phi\\ \max\limits_{j\in I_D^i}\{d_{ij}\}, & I_D^i=\phi\end{cases}。\tag{1}$$

式(1)含义为：当 x_i 具有最大局部密度时，δ_i 表示与 x_i 的距离最大的数据点与 x_i 之间的距离；否则，δ_i 表示在所有局部密度大于 x_i 的数据点中与 x_i 距离最小的数据点与 x_i 之间的距离。

1.2.3 精英点综合考察量

综合两种因素，定义一个将局部密度 ρ_i 和高密度点距离 δ_i 综合考虑的量

$$\gamma_i=\rho_i\delta_i,\quad i\in I_D。\tag{2}$$

显然 γ_i 值越大越有可能是精英点，因此只需对 $\{\gamma_i\}_{i=1}^{n}$ 进行降序排列，然后依序截取若干个数据点即可。设有样本有序集合 $D_\gamma=\{\gamma_{(1)},\gamma_{(2)},.\cdots,\gamma_{(n)}\}\cdots\cdots$，其中 $\gamma_{(i)}$ 表示降序排序位于第 i 个位置的值，$x_{(i)}$ 表示与 $\gamma_{(i)}$ 相对应的数据样本。给出精英保留比例 P_{elite}，可以通过公式(3)计算得到精英点集合：

$$D_{\text{elite}}=\{x_{(i)}\}_{i=1}^{\lceil n\times P_{\text{elite}}\rceil}。\tag{3}$$

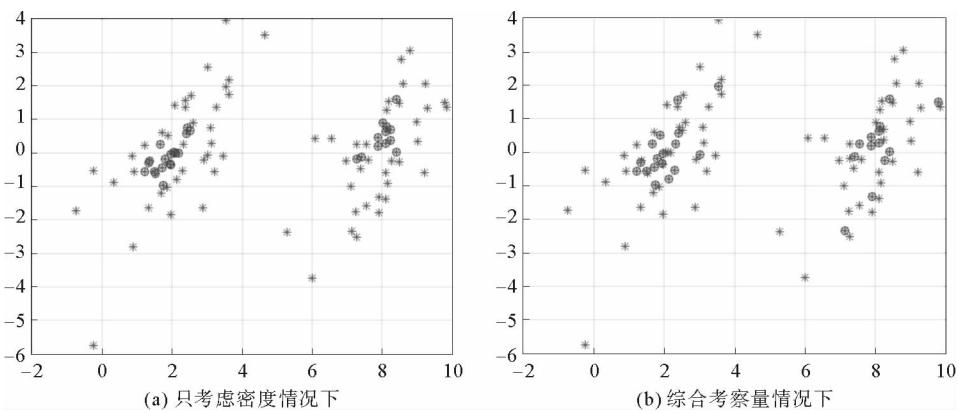

图 3 精英点选取情况对比

Fig. 3 Comparison of elite selection

根据给出的 γ 值标准，可以对样本数据进行选取。图 3 是保留比例为 30%情况下，精英点选取情况对比其中被圈出的点是精英点。

2 样本约简算法

基于精英点选取标准，提出了基于密度的 DBES(density-based elite selecting)样本约简算法如下：

输入：	样本集 D，阈值截取位置 P_{dc}，精英保留比例 P_{elite}
输出：	样本子集 D_{elite}
1.	计算得到距离矩阵 $\boldsymbol{D}_d$；
2.	根据 P_{dc} 得到对应位置的 d_c 及密度 ρ；
3.	for 样本集中的每一条数据 x_i
4.	If x_i 具有最大局部密度
5.	找到与 x_i 的距离最大的数据点 x_j；$\delta_i = d_{ij}$；
6.	else
7.	找到密度比 x_i 大的数据点中，与 x_i 距离最小的数据点 x_k；$\delta_i = d_{ik}$；
8.	end if
9.	end for
10.	计算 γ 值；根据 P_{elite} 值保留对应百分比的样本集 D_{elite}。

算法中，样本集 D 表示的是带约简的原始样本集；P_{dc} 决定了算法所得到密度的有效性，同时也会影响到选取样本子集的有效性；P_{elite} 决定了样本子集的大小，可以根据需要进行设定。关于 P_{dc} 和 P_{elite} 这两个参数的设定将在第 3.3 节进行讨论。

假设样本集中包含数据点个数为 n，计算距离矩阵 $\boldsymbol{D}_d$ 的时间复杂度为 $O(n^2)$。由于距离矩阵 $\boldsymbol{D}_d$ 是对称矩阵，所以在实际计算时可以计算一个上三角矩阵，节省一半的计算时间。计算 δ 值，需要比较两两之间的密度，时间复杂度为 $O(n^2)$。和计算距离矩阵 $\boldsymbol{D}_d$ 类似，根据矩阵的对称性可以节省一半的时间。算法中其他部分时间复杂度很低，因此整个算法的时间复杂度为 $O(n^2)$。

表 1 样本集说明

Tab. 1 Description of data set

名称	样本数	特征	数据来源
a1a	1 605	123	UCI/Adult
australian	690	14	Statlog/Australian
breast-cancer	683	10	UCI/Wisconsin Breast Cancer
fourclass	862	2	TKH96a[24]
german _ numer	1 000	24	Statlog/German
heart	270	13	Statlog/Heart
ionosphere	351	34	UCI/Ionosphere
mushrooms	8 124	112	UCI/mushrooms
sonar	208	60	UCI/Undocumented/Sonar
svmguide1	3 089	4	CWH03a[25]
w1a	2 477	300	JP98a[26]

3 实验分析与讨论

实验包括三部分。3.1 节是本文算法与其他样本约简算法的对比实验，对算法的约简比例和在分类算法上的分类精度进行了实验。其中用于对比的样本约简算法考虑到了所属类别以及影响力[11]，用于测试的分类算法包括 SVM、KNN 和 C4.5 三种不同类型的分类算法。3.2 节是样本约简后对分类的影响实验。3.3 节是算法约简稳定性及参数设置的分析实验，通过实验给出了 DBES 算法的参数设置建议。

实验所用样本集见表 1，表 2 是选择对比的样本约简算法。

3.1 约简算法实验对比

本节选择了几种不同类型的算法进行对比分析。表 3 是样本约简比例(reduction rate，RR)的对比。表

4～7是分类精度的对比，其中DBES[1]和DBES[2]分别表示：在P_{dc}取值分别为0.02的情况下，P_{elite}取值为0.3和取值为0.5时的计算结果。

从表3可以看出，MCS算法在w1a样本集上没有实现有效约简，样本集大小无变化；在mushrooms样本集上，MCS、ICF和DROP3算法得到的样本子集偏小，而RENN算法得到的样本子集和原样本集相同，约简效果较差；相比于其他算法，DBES算法可以对样本集进行有效约简，并且根据P_{elite}的不同，可以得到任意大小的样本子集。通过调整参数P_{elite}，可以使得约简比例高于其他的几种算法。

表2 选择对比的样本约简算法

Tab.2 Comparison of selected instance reduction algorithm

算法名称	算法类型	算法时间复杂度
MCS	能力保持型	$O(n^2)$
ICF	混合型	$O(n^2)$
DROP3	混合型	$O(n^2)$
RENN	能力增强型	$O(n^2)$

表3 样本约简比例对比

Tab.3 Comparison of instance reduction ratio

样本集	MCS	ICF	DROP3	RENN	DBES1	DBES2
a1a	**0.00**	0.25	0.25	0.26	0.70	0.50
australian	0.00	0.81	0.84	0.79	0.70	0.50
breast-cancer	0.92	0.79	0.95	**0.04**	0.70	0.50
fourclass	0.41	0.82	0.89	0.60	0.70	0.50
german_numer	**0.00**	0.96	0.98	0.55	0.70	0.50
heart	0.65	0.64	0.64	0.25	0.70	0.50
ionosphere	**0.00**	0.72	0.77	**0.00**	0.70	0.50
mushrooms	0.99	0.94	0.95	**0.00**	0.70	0.50
sonar	0.74	0.52	0.52	0.17	0.70	0.50
svmguide1	0.92	0.76	0.81	0.05	0.70	0.50
w1a	**0.00**	**0.03**	**0.03**	**0.05**	0.70	0.50
average	0.42	0.66	0.69	0.25	0.70	0.50

表4 样本子集分类精度对比(SVM)

Tab.4 Classification accuracy comparison (SVM)

样本集	MCS	ICF	DROP3	RENN	DBES1	DBES2
a1a	0.817	0.754	0.755	0.819	0.815	0.819
australian	0.858	0.866	0.831	0.551	0.855	0.855
breast-cancer	0.973	0.954	0.903	0.972	0.973	0.973
fourclass	0.799	0.777	0.712	0.752	0.799	0.808
german_numer	0.811	0.763	0.775	0.700	0.755	0.785
heart	0.851	0.774	0.855	0.848	0.792	0.818
ionosphere	0.945	0.367	0.601	0.945	0.772	0.903
mushrooms	0.518	0.505	0.505	0.999	0.739	0.895
sonar	0.514	0.701	0.860	0.812	0.711	0.750
svmguide1	0.400	0.935	0.864	0.974	0.646	0.646
w1a	0.970	0.971	0.971	0.970	0.970	0.970
average	0.769	0.761	0.785	0.850	0.803	0.839

表4是样本子集的分类精度对比，使用约简后的样本子集作为训练集，用LIBSVM进行训练得到分类模型，在测试集上进行测试得到分类精度。表格中average行表示样本约简比例以及样本子集分类精度的均值。表5和表6分别是KNN算法和C4.5算法的实验结果。结合表3、表4和表5可以看出，采用不同的分类算法测试，DBES算法都有较好的表现，其中在决策树(C4.5)算法下的优势最为明显。由于提出的准则以及算法不针对某种分类器，因此受分类算法的影响较小；与之相比较，其他算法受分类算法的影响较大。另外，对比3个表中DBES算法在P_{elite}不同取值时的分类精度可以看出，增大P_{elite}可以提高DBES算法的分类精度。

通过表3～6中的average可以看到，有的算法对样本集平均约简比例较大，但分类精度偏低，如DROP3算法；有的算法分类精度较高，但对样本集平均约简比例较小，如RENN算法；相比较于其他的方法，DBES算法能够在两个方面都保持较好的效果。

为了更好的说明DBES算法的优越性，采用文献[11]中的约简算法比较方法，将约简比例和分类精度的乘积作为新的综合指标，得到约简效果是综合指标对比如表7，从表中可以看出DBES算法($P_{elite}=0.3$)的综合指标最高且非常稳定。

表8和图4是各个算法在不同样本集上分类精度和约简比例的方差。方差的大小反映了算法的稳定性。DBES算法可以根据参数将样本集约简到指定大小，约简稳定性高；对比不同分类算法上的分类精度方

差可以发现，本文提出算法方差都非常小，分类精度的稳定性也明显好于其他算法。

3.2 约简算法对分类的影响

本节以 SVM 为例，采用交叉验证的方式，对算法进行了测试。实验选择了 fourclass、heart 和 a1a 样本集。其中 fourclass 是人工数据集，heart 是低维真实数据集，a1a 是高维真实数据集。实验结果如图 5～7 所示，图中实线代表约简后的分类精度及其均值，虚线代表原样本集分类精度及其均值。

表 5 样本子集分类精度对比(KNN)

Tab. 5 Classification accuracy comparison (KNN)

样本集	MCS	ICF	DROP3	RENN	DBES1	DBES2
a1a	0.883	0.753	0.753	0.745	0.796	0.809
australian	0.905	0.766	0.788	0.872	0.853	0.873
breast-cancer	0.959	0.847	0.937	0.880	0.973	0.975
fourclass	1.000	0.849	0.920	1.000	1.000	1.000
german_numer	0.836	0.605	0.651	0.760	0.708	0.759
heart	0.796	0.722	0.711	0.779	0.800	0.811
ionosphere	0.860	0.359	0.359	0.862	0.826	0.906
mushrooms	0.943	0.482	0.482	0.816	0.859	0.898
sonar	0.729	0.615	0.649	0.740	0.725	0.754
svmguide1	0.958	0.631	0.917	0.943	0.956	0.959
w1a	0.986	0.970	0.970	0.918	0.890	0.894
average	0.896	0.691	0.740	0.847	0.853	0.876

表 6 样本子集分类精度对比(C4.5)

Tab. 6 Classification accuracy comparison (C4.5)

样本集	MCS	ICF	DROP3	RENN	DBES1	DBES2
a1a	0.940	0.753	0.753	0.794	0.810	0.839
australian	0.713	0.879	0.871	0.833	0.847	0.856
breast-cancer	0.803	0.884	0.940	0.612	0.935	0.945
fourclass	0.720	0.768	0.660	0.736	0.938	0.967
german_numer	0.774	0.724	0.745	0.754	0.752	0.806
heart	0.711	0.788	0.818	0.768	0.755	0.770
ionosphere	0.813	0.359	0.359	0.768	0.755	0.846
mushrooms	0.972	0.482	0.482	0.950	0.871	0.912
sonar	0.740 4	0.721	0.745	0.705	0.635	0.749
svmguide1	0.938	0.933	0.953	0.928	0.948	0.964
w1a	0.987	0.970	0.970	0.940	0.952	0.981
average	0.828	0.751	0.754	0.799	0.836	**0.876**

表 7 约简效果综合指标对比

Tab. 7 Comparison of comprehensive index of reduction

	MCS	ICF	DROP3	RENN	DBES1	DBES2
SVM	0.276	0.486	0.527	0.179	**0.562**	0.419
KNN	0.378	0.432	0.499	0.213	**0.597**	0.438
C4.5	0.351	0.477	0.507	0.195	**0.585**	0.438
average	0.335	0.465	0.511	0.196	**0.581**	0.432

表 8 约简比例和分类精度的方差

Tab. 8 Variance of reduction ratio and classification accuracy

	MCS	ICF	DROP3	RENN	DBES1	DBES2
RR	0.186	0.083	0.096	0.076	**0.000**	**0.000**
SVM	0.039	0.034	0.018	0.020	**0.010**	**0.009**
KNN	0.007	0.030	0.038	0.007	**0.009**	**0.007**
C4.5	0.012	0.034	0.037	0.011	**0.010**	**0.006**

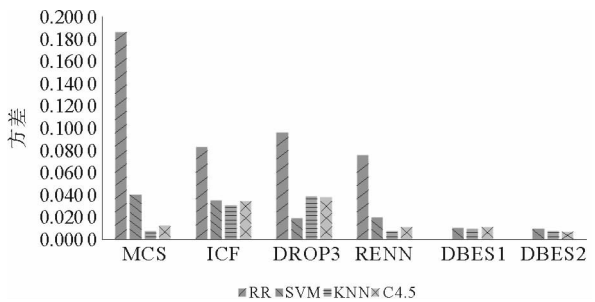

图 4 约简比例和分类精度的方差

Fig. 4 Variance of reduction ratio and classification accuracy

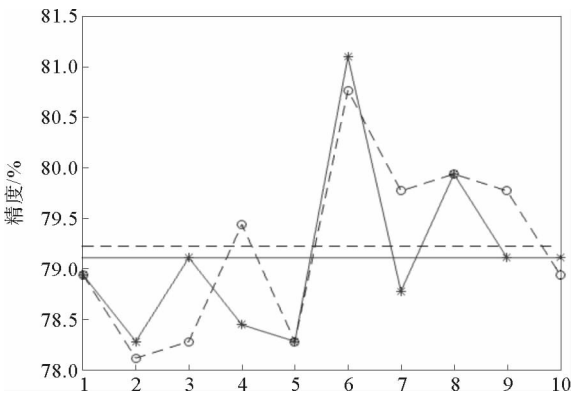

图 5 fourclass 随机抽样验证结果

Fig. 5 Random sample validation results of data set 1

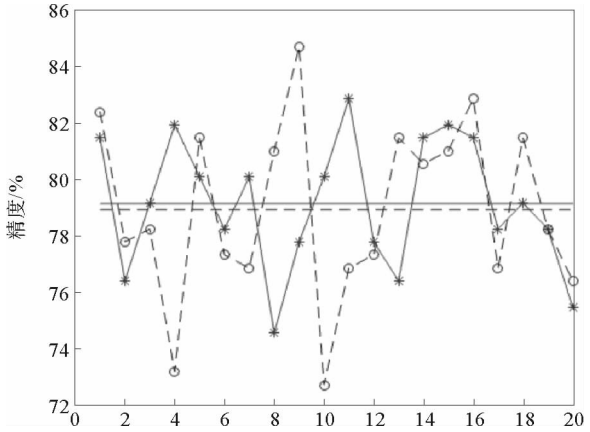

图 6　heart 随机抽样实验结果

Fig. 6　Random sample validation results of data set 2

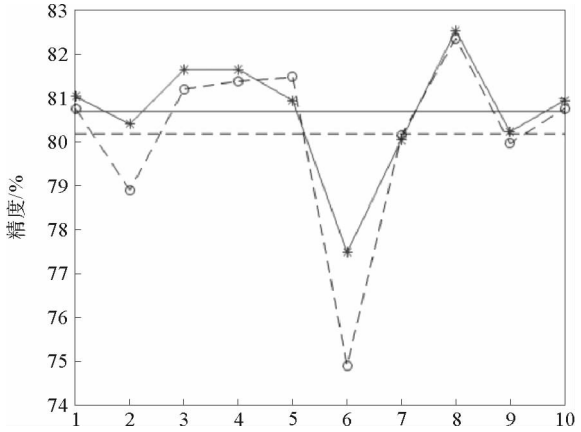

图 7　a1a 降维后的实验结果

Fig. 7　Results of data set 3 after dimension reduction

在 fourclass 上进行试验，随机抽样 10 次，抽取比例为 50%，阈值截取位置 $P_{dc}=0.2$，精英保留比例 $P_{elite}=0.25$。由图 5 可以看出，因为 fourclass 为人工样本集合，样本点是使用算法按规则产生的，随机抽样验证结果的均值非常接近，其中有些情况下的精度甚至超过了原样本集的结果。

在 heart 上进行试验，随机抽样 20 次，抽取比例为 50%，阈值截取位置 $P_{dc}=0.2$，精英保留比例 $P_{elite}=0.3$。由图 6 可以看出，该算法在该样本集上有较好的表现，约简后样本子集的分类结果甚至超过了原样本集的结果。真实数据中存在噪声数据，使用这样的数据进行训练，分类模型很可能受到噪声数据的影响。提出的约简算法能够过滤掉噪声数据，从而提高分类精度。

DBES 算法在处理高维数据时效果有所下降。由于高维数据的稀疏性，将低维空间中的距离度量函数应用到高维空间时，随着维数的增加，数据对象之间距离的差异将不复存在，其有效性大大降低。实际应用中最常用的是 PCA 降维的方法。图 7 是对 a1a 使用 PCA 降到 10 维之后的实验结果，随机抽样 10 次，抽取比例为 50%，阈值截取位置 $P_{dc}=0.02$，精英保留比例 $P_{elite}=0.3$。

通过以上实验可以看出，提出的样本约简算法能够较好地保持原分类算法的分类精度；对某些包含噪声的样本集合能够提高分类精度；对于高维数据经过降维处理之后，也能保持分类精度。

3.3　算法中参数的讨论

算法中包含三个参数，除了参数 D，其他两个参数都需要给定。其中 P_{dc} 决定了算法所得到密度的有效性，P_{elite} 决定了样本子集的大小。P_{dc} 和 P_{elite} 的取值都在(0,1]之间并且由用户给定。图 8 表示的是 $P_{elite}=0.3$ 时，P_{dc} 取值变化对分类精度的影响(其中横轴代表 P_{dc} 取值变化，纵轴代表分类精度百分比)。图 9 表示的是 $P_{dc}=0.2$ 时，P_{elite} 取值变化对分类精度的影响(其中横轴代表 P_{elite} 取值变化，纵轴代表分类精度百分比)。

从图 8 中可以看出，P_{dc} 的取值在[0.1,0.4]区间有较稳定的表现，取值过大或过小都会影响算法的稳定性；虽然在某些样本集上如 ionosphere，随着 P_{dc} 的增大，分类精度有所提升，但在大部分样本集上，都会保持不变或者下降。从图 9 中可以看出，当 P_{elite} 的取值小于 0.2 时，分类精度会出现较大的波动。结合实际分析，保留更多的样本点意味着保留了更多的信息，会提高分类的精度。P_{elite} 的增大意味着保留的样本子集逐渐增大，当 P_{elite} 值为 1 时意味着没有约简。所以，P_{elite} 的取值可以根据样本数据和实际需求自行设定，建议取值区间为[0.2,1]。

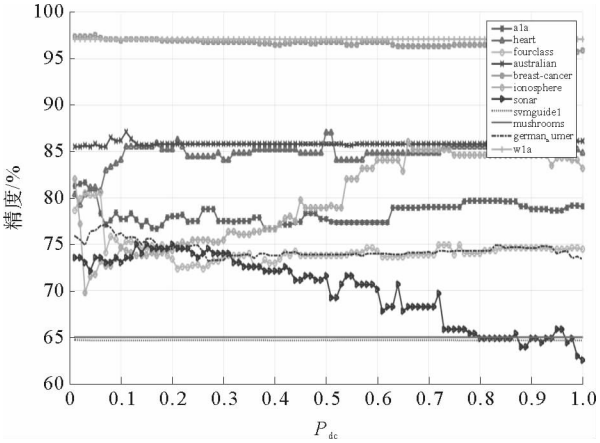

图 8　P_{dc} 取值变化对分类精度的影响

Fig. 8　Effect of change of P_{dc} on classification accuracy

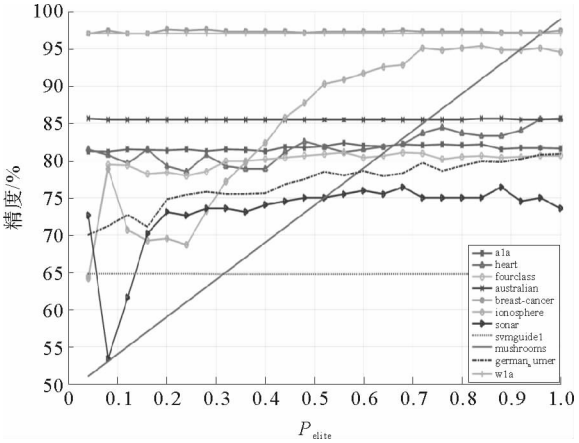

图 9　P_{elite} 取值变化对分类精度的影响

Fig. 9　Effect of change of P_{elite} on classification accuracy

4　结语

本文给出了一种新的样本点选取标准，并基于该标准设计实现了一种适用于不同分类器的样本约简算法。与其他算法相比较，该算法可以根据需求获得任意大小的样本子集。与多种分类算法进行测试对比，算法的约简效果综合指标要优于其他算法。对包含噪声点的样本集，算法的分类精度和稳定性均有一定程度的提高。在模式挖掘中使用该算法，可以提高模式挖掘效率。如何针对大数据做并行化改进，是下一步需要研究的问题。

参考文献：

[1]WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems Man & Cybernetics，1972，2(3)：408-421.

[2]WILSON D R，MARTINEZ T R. Reduction techniques for instance-based learning algorithms[J]. Machine Learning，2000，38(3)：257-286.

[3]TOMEK I. An experiment with the edited nearest-neighbor rule[J]. IEEE Transactions on Systems Man & Cybernetics，1976，SMC-6(6)：448-452.

[4]HART P. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory，1968，14(3)：515-516.

[5]GATES G. The reduced nearest neighbor rule[J]. IEEE Transations on Information Theory，1972，18(3)：431-433.

[6]DASARATHY B V. Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design[J]. IEEE Transactions on Systems Man & Cybernetics，1994，24(3)：511-517.

[7]ANGIULLI F. Fast nearest neighbor condensation for large data sets classification[J]. IEEE Transactions on Knowledge and Data Engineering，2007，19(11)：1450-1464.

[8]BRIGHTON H，MELLISH C. Advances in instance selection for instance-based learning algorithms[J]. Data Mining & Knowledge Discovery，2002，6(2)：153-172.

[9]WILSON D R，MARTINEZ T R. Reduction techniques for instance-based learning algorithms[J]. Machine Learning，2000，38(3)：257-286.

[10]REINARTZ T. A unifying view on instance selection[J]. Data Mining & Knowledge Discovery，2002，6(2)：191-210.

[11]GARCIA S，DERRAC J，CANO J R，et al. Prototype selection for nearest neighbor classification：Taxonomy and empirical study[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2012，34(3)：417-435.

[12]CHEN J,ZHANG C,XUE X,et al. Fast instance selection for speeding up support vector machines[J]. Knowledge-Based Systems,2013,45(3):1-7.

[13]WANG X Z,DONG L C,YAN J H. Maximum ambiguity-based sample selection in fuzzy decision tree induction[J]. IEEE Transactions on Knowledge & Data Engineering,2012,24(8):1491-1505.

[14]HAMIDZADEH J,MONSEFI R,YAZDI H S. LMIRA:Large margin instance reduction algorithm[J]. Neurocomputing,2014,145(18):477-487.

[15]HAMIDZADEH J,MONSEFI R,YAZDI H S. IRAHC:Instance reduction algorithm using hyperrectangle[J]. Pattern Recognition,2015,48(5):1878-1889.

[16]NIKOLAIDIS K,RODRIGUEZ-MARTINEZ E,GOULERMAS J Y,et al. Spectral graph optimization for instance reduction[J]. IEEE Transactions on Neural Networks & Learning Systems,2012,23(7):1169-1175.

[17]GARCIA-LIMON M,ESCALANTE H J,MORALES-REYES A. In defense of online Kmeans for prototype generation and instance reduction[C]//ACM Conference on Computer Supported Cooperative work,CSCW 2004,Chicago,2016:274-283.

[18]JAMJOOM M,HINDI K E. Partial instance reduction for noise elimination ☆[J]. Pattern Recognition Letters,2016,74:30-37.

[19]RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344 (6191):1492-1496.

[20]LEWIS D D,GALE W A. A sequential algorithm for training text classifiers[C]// Acm Sigir Forum. Springer-Verlag New York,Inc. 1994:3-12.

[21]SETTLES B. Active learning literature survey[J]. University of Wisconsinmadison,2010,39(2):127-131.

[22]DAGAN I,ENGELSON S P. Committee-based sampling for training probabilistic classifiers[J]. Machine Learning Proceedings,1995:150-157.

[23]SUYKENS J A K,VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters,1999,9(3):293-300.

[24]THO T K,KLEINBERG E M. Building projectable classifiers of arbitrary complexity[C]// International Conference on Pattern Recognition. IEEE Computer Society,1996:880.

[25]HSU C W,CHANG C C,LIN C J. A practical guide to support vector classification[R]. Taiwan: National Taiwan University.

[26]PLATT J C. Fast training of support vector machines using sequential minimal optimization[C]// Advances in Kernel Methods-support Vector Rning. MIT Press,1999:185-208.

(责任编辑:傅　游)