基于稀疏表示的不平衡数据集过采样算法

訾壮壮[1] 何 涛[2] 赵 停[1]

[1](南京邮电大学电子与光学工程学院 江苏 南京 210023)

[2](南京邮电大学工程训练中心 江苏 南京 210023)

摘 要 大多数不平衡数据集过采样方法依赖于欧几里得特征空间中少数类样本的空间位置,使用少数类样本的局部信息生成新样本来减轻类不平衡问题,因此新生成的少数类样本质量较差。针对这种情况,提出一种 K 稀疏解过采样算法(K Sparse Over-Sampling,KSOS),其使用少数类样本的全局信息进行样本合成。使用少数类样本来构造稀疏字典,通过求解 L_1 范数最小化来获得当前点的稀疏解;使用稀疏解中的非零项所对应的项来生成新的样本;计算每一个新生成样本的置信度,将所有新生成样本按其置信度排序,从中选取符合要求的新生成样本。在几个 UCI 数据集上的实验结果证明了该算法的有效性。

关键词 不平衡数据集 过采样 K 稀疏过采样

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.10.046

SPARSE REPRESENTATION-BASED OVER-SAMPLING ALGORITHM OF IMBALANCED DATASET

Zi Zhuangzhuang[1] He Tao[2] Zhao Ting[1]

[1](*College of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, Jiangsu, China*)

[2](*Engineering Training Center, Nanjing University of Posts and Telecommunications, Nanjing 210023, Jiangsu, China*)

Abstract Most over-sampling methods of imbalanced dataset depend on the spatial location of minority samples in Euclidean feature space, and use the local information of minority samples to generate new samples to alleviate the problem of class imbalance. Therefore, the quality of the newly generated minority class samples is poor. In view of this, we propose a K-sparse solution over-sampling algorithm, KSOS, which uses global information of minority class samples for sample synthesis. It used the minority class samples to construct the sparse dictionary, and obtained the sparse solution of the current point by solving the L_1 norm minimization; the non-zero term in the sparse solution was used to generate new samples; the confidence level of each newly generated sample was calculated in the sample validation phase; all the newly generated samples were sorted according to their confidence, and the new generated samples that meet the requirements were selected from them. The experimental results on several UCI datasets show the effectiveness of the KSOS algorithm.

Keywords Imbalanced dataset Over-sampling KSOS

0 引 言

不平衡数据作为数据挖掘中最具挑战性的问题之一[1],越来越受到人们的重视。不平衡数据是指不同类别样本数量存在比例失衡的数据集。尽管大多数据具有多类属性,但是许多情况下仍然可以转换为二分类问题,因此本文主要研究二分类问题。在只有两个类别的数据集中,多数类是指样本数量相对较多的类,少数类是指样本数量相对较少的类。不平衡数据集在欺诈检测、文本分类、医疗诊断、推荐系统和其他实际应用中非常普遍。在上述应用中,人们通常对少数类

收稿日期:2019-05-19。**訾壮壮**,硕士生,主研领域:不平衡数据分类。**何涛**,讲师。**赵停**,硕士生。

样本更感兴趣,然而传统的分类器通常偏向多数类,无法对少数类样本进行正确分类,因此性能较差。

有两种方法可以用于解决类不平衡问题,它们分别是数据层面的解决方案和算法层面的解决方案。数据层面的解决方案是通过欠采样和过采样或二者的组合来平衡少数类样本和多数类样本的分布。算法层面的解决方案是通过改进传统的分类算法或优化学习算法的性能来提高少数类样本的识别率[2]。数据层面解决方案由于其独立于分类算法,可与任何传统的分类算法相结合而更受欢迎。

大多数不平衡数据集过采样方法通过生成新的少数类样本以减轻类不平衡问题,这些方法通常依赖于欧几里得特征空间中少数类样本的空间位置,收集少数类样本的局部信息以生成新的少数类样本。使得新生成的样本不遵循原始少数类样本的分布,因此含有较少的信息量。而更好的想法是利用少数类样本的全局信息,这可以在生成新样本的同时考虑稀疏表示来实现。

本文提出一种基于稀疏表示的不平衡数据集过采样算法 KSOS。在样本生成阶段,使用少数类样本来构造稀疏字典,通过求解 L_1 范数最小化来获得当前点的稀疏解,然后使用稀疏解中的非零项所对应的少数类样本来生成新样本。在样本确认阶段,计算每一个新生成样本的置信度,然后将所有新生成样本按其置信度排序,从中选取符合要求的新生成样本。

1　相关工作

不平衡主要有三种:类间不平衡、类内不平衡和类重叠。几乎所有的过采样方法都可以通过重复原始少数类样本或生成新的少数类样本来解决类间不平衡问题。对于类内不平衡问题,常见方法是首先使用聚类技术来发现少数类子群,然后应用过采样方法来处理这些子群,或者在用其局部分布衡量学习难度后,对难以学习的少数群体样本进行加权。而当原始非重叠区域被噪声和错误的生成样本入侵时,就会出现类重叠问题。

随机过采样通过简单重复部分少数类样本的方式来达成少数类和多数类样本在训练规模上的平衡。然而随机过采样技术可能导致过拟合问题,因为它可能创建比原始数据集更小且更具特异性的决策区域。

使用基于样本生成的过采样技术来解决类不平衡问题也很普遍。SMOTE[3] 被提出用于生成少数类样本,以扩大少数类样本的原始决策区域。SMOTE 随机选择一个少数类样本 x,使用 KNN 在其余所有少数类样本中选出 x 的 K 个近邻样本,并取出其中任意一个 x',然后在 x 和 x' 之间连线的任一位置生成新的少数类样本。SMOTE 虽然在减少过拟合方面取得了进展,但却引发了过度泛化的问题。许多 SMOTE 的扩展算法也相继被提出,例如 B-SMOTE[4]、Safe-Level-SMOTE[5]、Random-SMOTE[6],实验表明生成的样本比简单复制的样本更具信息量。

自适应技术也被广泛用于不平衡数据集过采样。自适应合成采样 ADASYN[7] 考虑到难以学习的少数类样本,并根据其局部分布自动合成少数类样本。自适应半无监督加权过采样 A-SUWO[8],应用半无监督分层聚类方法对少数类样本进行聚类,并根据其特定大小自适应地对每个子聚类进行过采样。

2　算法设计

2.1　压缩感知和稀疏表示

压缩感知(也称为压缩感测、压缩采样或稀疏采样)是通过寻找欠定线性系统的解,来有效地获取和重建信号的信号处理技术。奈奎斯特-香农采样定理表明:如果信号 $x(t)$ 不包含高于 B 赫兹的频率,那么其可以通过一系列间隔 1/2B 的点处的值来完全确定。过去几十年中该定理一直被应用于数字信号处理。然而压缩感知定理指出信号带宽不是采样的基本要求,信号采样率仅取决于采样系统的稀疏性和非相干性,该理论可以同时实现信号的压缩和采样,在学术界和工业界得到了广泛的应用。

压缩感知的核心问题是稀疏字典和测量矩阵的设计,以及信号重构算法的构造,其中信号重构算法的构造也被称为稀疏表示。

观测数据 $\boldsymbol{y}$ 是长度为 M 的列向量,即 $\boldsymbol{y} \in \mathbf{R}^{M\times 1}$,采样信号 $\boldsymbol{A} \in \mathbf{R}^{M\times N}(M \ll N)$ 为一组基向量。压缩感知的目标是使用线性联立方程 $\boldsymbol{y} = \boldsymbol{A}x$ 从观测数据 $\boldsymbol{y}$ 中恢复被测信号 x。由于未知数的数量大于方程组的个数使得欠定方程组是病态的,它的解不存在。

如果使被测信号 x 变得稀疏,可能意味着 $\| x \|_0$ (x 的 L_0 范数)尽可能小,那么未知数的数量将会显著下降,这使得信号重建成为可能。

因此,压缩感知问题归结为求解如下约束优化问题:

$$\min \| x \|_0 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}x \tag{1}$$

2005 年,Starck 等[11] 证明式(1)具有唯一解,但是 L_0 范数最小化是非凸优化问题。由于在多项式时间内不能获得可行解,因此式(1)的求解是个 NP 难问

题。2006 年,Tsaig 等[12]证明 L_1 范数最小化可以基于 RIP 条件替代 L_0 范数最小化[12]。尽管两种形式都具有相同的稀疏解,但压缩感知的框架变为凸优化问题,其优化目标如下:

$$\min \| x \|_1 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}x \tag{2}$$

使用它,形成了压缩感知最初的框架。实际上,当考虑到噪声问题时我们通常使用式(3)而不是式(2)。

$$\min \| x \|_1 \quad \text{s.t.} \quad \| \boldsymbol{A}x - \boldsymbol{y} \|_2 \leqslant \varepsilon \tag{3}$$

式中:ε 表示大于 0 的很小的数;$\boldsymbol{A}$ 表示稀疏字典;x 为稀疏解。

重建算法的目的是找到解 x,其中整个问题的核心是 $\boldsymbol{y}$ 的稀疏表示。

2.2 KSOS

稀疏字典的构造包括人工构造和训练学习。前者包含各向同性 Gabor 字典、各向异性精化-高斯字典等;后者包含字典学习算法 K-SVD。本文直接采用训练样本构造稀疏字典。

首先将所有少数类 $S_{\min}$ 从训练集 S 中分离出来。其中 $S_{\min} \in \mathbf{R}^{m \times n}$,$m$ 为少数类样本数目,n 为少数类样本的维度。对于当前采样点 x_i,$x_i \in S_{\min}$,使用除 x_i 之外的其余少数类样本来构造系数字典 D,$D \in \mathbf{R}^{n \times (m-1)}$。

接着,按照式(4)对每个样本进行标准化并计算它们的 L_2 范数。

$$y'_{i,j} = \frac{y_{i,j}}{\sqrt{\sum_j y_{i,j}^2}}$$
$$i = 1,2,\cdots,m-1 \quad j = 1,2,\cdots,n \tag{4}$$

式中:$y_{i,j}$ 是稀疏字典 D 的样本点。在得到稀疏 D 之后,可以通过求解 L_1 最小化问题来求解少数类样本的稀疏解 w。

优化目标如下:

$$\hat{W}_1 = \operatorname{argmin} \| w \|_1 \tag{5}$$
$$\text{s.t.} \quad \| Dw - x \|_2 \leqslant \varepsilon$$

通过使用同伦算法[9]来求解方程的解。为了降低噪声,同伦算法考虑以下基本问题:

$$\min_{x,\lambda} \frac{1}{2} \| x - Dw \|_2^2 + \lambda \| w \|_1 \tag{6}$$

式中:λ 是拉格朗日乘子。

在获得稀疏解后,选择 w 中非零项标识的样本点与 x_i 一起合成新样本。例如,得到 x_i 的非零稀疏解的下标,并将其索引保存在 $\boldsymbol{A}$ 中。然后选择 $\boldsymbol{A}$ 中所有的元素,它代表数据点 y_k 的下标,其中 k 代表 $\boldsymbol{A}$ 中元素个数。通过式(7)合成一个新样本。

$$x_{\text{new}} = x_{i,l} + \sum_{j=1}^{k} (y_{j,l} - x_{i,l})\delta_j \tag{7}$$
$$i = 1,2,\cdots,M \quad l = 1,2,\cdots,n$$

式中:

$$\delta \in [0,1]$$
$$\delta_1 \in [0,1-\delta]$$
$$\delta_2 \in [0,1-\delta-\delta_1]$$
$$\vdots$$
$$\delta_k \in [0,1-\delta-\delta_1-\cdots-\delta_{k-1}]$$

M 为特征数,n 为样本维度。

在样本确认阶段,KNN 模型中定义了一个样本的置信度,这表明样本最近邻的分布。样本的置信度定义如下:

$$Confidence(T) = \frac{M}{T} \tag{8}$$

式中:T 是合成少数类样本最近邻的总数;M 是 K 近邻中属于少数样本的数目。例如,如果一个样本的 5 个最近邻中的 4 个属于少数类,那么该样本的置信度为 3.2。然后将所有新生成样本按置信度从大到小排序并从中选取样本。

算法 1 KSOS 算法

输入:训练集 $S = \{(x_i,z_i), i = 1,2,\cdots,N, z_i \in \{0,1\}\}$;多数类样本 N^-,少数类样本 N^+,其中,$N^+ + N^- = N$;采样率 $SR\%$。

输出:过采样后的训练集 $S' = \{(x_i,z_i), i = 1,2,\cdots,N + N^+ \times SR, z_i \in \{0,1\}\}$。

1. 从训练集 S 中取出全部多数类样本与少数类样本,组成多数类训练样本集及少数类训练样本集 S^+
2. 置新生成样本集 S^{new} 为空
3. for $i = 1:N^+$
4. 用除 x_i 外的所有少数类训练样本构建稀疏字典 D,其中 $x_i \in S^+$
5. 用同伦算法解决式(5)所示的 L_1 优化问题得出非零稀疏解所对应 D 中的少数类样本点,并将其置于稀疏解对应样本集 S^{new} 中
6. for $i = (SR/100) \times 2$
7. 在稀疏解对应样本集 S^{new} 中取出对应的少数类样本点 y_j,$j = 1,2,\cdots,k$,其中 k 为 x_i 对应非零稀疏解的个数
8. $x_{\text{new}} = x_{i,l} + \sum_{j=1}^{k} (y_{j,l} - x_{i,l})\delta_j$
9. 添加 $x_{\text{new};k}$ 至 S^{new},$S^{\text{new}} = S^{\text{new}} \cup x_{\text{new}}$
10. 置 S^{ner} 为空
11. end for
12. end for
13. 计算每个新生成样本的置信度,并将其从大到小排列,从中选取前 $N^+ \times (SR/100)$ 个置于 S^{new} 中
14. 得到过采样后的训练集 $S' = S \cup S^{\text{new}}$

为对比 KSOS 算法与 SMOTE 和 ADASYN 的采样

效果,在一个合成数据集上进行实验。为了方便可视化,该合成数据集的维度为 2,其中少数类样本 20 个,多数类样本 200 个。少数类样本以(2.5,2.5)为中心,分别以 0.1 和 0.4 作为第一维和第二维数据的标准差生成的高斯随机变量,多数类样本以(0.3, 0.3) 为中心,分别以 0.2 和 0.2 作为第一维和第二维数据的标准差生成的高斯随机变量。图 1 为在合成数据集上三种采样算法的采样结果。可以看出,ADASYN 和 SMOTE 都一定程度上在位于多数类样本区域的少数类样本附近生成了新样本,造成噪声信息的传播。而 KSOS 利用少数类样本的全局信息进行样本生成,新生成的样本多位于少数类样本及其边缘,不仅能改善噪声信息的传播问题,也增强了少数类样本的决策边界。

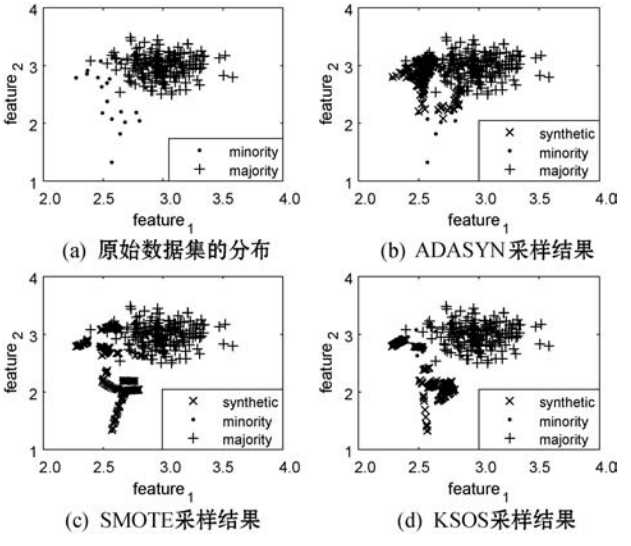

图 1　三种采样算法的效果对比图

3　实　验

3.1　数据集描述

为了测试本文提出方法的分类性能,实验采用 4 组 KEEL[1](表 1 前 4 组)和 2 组 HDDT[2](表 1 后 2 组)类别不平衡数据集。这些数据集的不平衡比率范围从 9.29 到 28.1。表 1 列出了数据集的基本信息,每个数据集有 6 个属性,即数据集的名称、数据集的特征维度、数据集的大小、少数类和多数类样数以及不平衡率。本文仅考虑含两个类别的数据集,因此需要对含有多类属性的数据集进行转换,以得到二分类标签。根据文献[10]中的方法来修改数据标签,即选取其中某一类作为少数类,并将其余所有的类合并为多数类。

表 1　不平衡数据集的基本信息

数据集	特征	数据集大小	少数类样本数	多数类样本数	不平衡率
Yeast4	8	1 484	51	1 433	28.10
Ecoli4	7	336	20	316	15.80
Ecoli0147VS56	6	332	25	307	12.28
Glass0146VS2	9	205	17	188	11.06
Covtype	10	38 500	2 746	35 754	13.02
satimage	36	6 430	625	5 805	9.29

3.2　评价指标

F-Measure:用于分类器查准率和召回率的性能度量,计算公式如下:

$$F\text{-}Measure = \sqrt{\frac{2 \times recall \times precision}{recall + precision}} \tag{9}$$

式中:*precision* 表示查准率;*recall* 表示召回率。

G-Mean:用于度量分类器在两类数据上的平均性能,计算公式如下:

$$G\text{-}Mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{10}$$

式中:*TP* 是实际为少数类且预测正确的样本个数;*TN* 是实际为多数类且预测正确的样本数量;*FN* 是实际为少数类且预测错误的样本数量;*FP* 是实际为多数类且预测错误的样本数量。

AUC:用于衡量分类器性能的指标,计算公式如下:

$$AUC = 1 - \frac{\frac{FP}{N} + \frac{FN}{M}}{2} \tag{11}$$

式中:*M* 和 *N* 分别表示数据集中少数类样本个数和多数类样本个数。

3.3　结果分析

在实验中使用 C4.5 决策树作为基础学习模型。为了考察本文算法的分类性能,将其与 SMOTE 和 ADASYN 过采样算法进行对比,作为参考,给出基于原始的不平衡数据集的 C4.5 决策树的学习性能。为保证对比算法的性能,其 K 近邻的参数设置与原文保持一致,即 SMOTE 和 ADASYN 的 K 近邻参数设置为 5,本文算法在样本确认阶段的 K 近邻参数设置为 3。对于每一个数据集,均采用 5 折交叉验证,每次选取其中 4 组作为训练集,1 组作为测试集,且每组数据中多数类与少数类样本个数的比值,为原始数据集中多数类

与少数类样本数量的不平衡比率。采用 G-Mean、F-Measure 和 AUC 作为评估指标,为尽量保证实验结果不受随机因素的干扰与影响,实验结果取 100 次 5 折交叉验证的均值,并将每个评价指标下的最大值与本文算法的获胜次数用粗体标出,如表 2、表 3 所示。

表 2 以 C4.5 决策树为分类器的 G-Mean、F-Measure 和 AUC 结果

数据集	方法	G-Mean	F-Measure	AUC
Yeast4	C4.5	0	0	0.458 505
	SMOTE	0.712 819	0.264 318	0.740 662
	ADASYN	0.676 513	0.247 470	0.702 914
	KSOS	**0.712 901**	**0.325 654**	**0.752 890**
Ecoli4	C4.5	0.803 135	0.641 154	0.939 236
	SMOTE	0.864 021	0.719 072	0.942 281
	ADASYN	0.849 413	0.705 175	0.942 485
	KSOS	**0.881 899**	**0.723 607**	**0.943 717**
Ecoli0147VS56	C4.5	0.756 112	0.647 778	0.989 572
	SMOTE	0.830 896	0.658 257	0.987 959
	ADASYN	0.763 055	0.556 737	0.987 087
	KSOS	**0.861 283**	**0.762 255**	**0.994 661**
Glass0146VS2	C4.5	0.194 500	0.116 667	0.962 731
	SMOTE	0.461 136	0.125 000	0.946 297
	ADASYN	**0.527 746**	1.833 260	0.952 762
	KSOS	0.507 609	**0.249 619**	**0.964 689**
Satimage	C4.5	0.694 965	0.560 860	0.776 062
	SMOTE	**0.823 932**	0.557 508	**0.851 473**
	ADASYN	0.817 673	0.487 830	0.844 813
	KSOS	0.778 984	**0.597 750**	0.825 635
Covtype	C4.5	0.627 782	0.459 244	0.681 703
	SMOTE	0.867 146	0.509 495	0.862 849
	ADASYN	0.845 001	0.541 765	0.844 450
	KSOS	**0.867 353**	**0.568 928**	**0.865 004**

表 3 算法获胜次数

	方法	G-Mean	F-Measure	AUC
获胜次数	C4.5	0	0	0
	SMOTE	1	0	1
	ADASYN	1	0	0
	KSOS	**4**	**6**	**5**

从各个算法在不同指标下的获胜次数可以看出,本文的方法在 AUC、G-Mean 和 F-Measure 三种评价指标下的性能都优于其他两种算法。其中 F-Measure 值在 6 个数据集上全都取得了最大值,而只有当查准率和查全率都比较高时,F-Measure 值才比较高,说明 KSOS 算法在不损害多数类数据分类性能的情况下,对少数类样本的查准率和查全率都有所提升。

4 结 语

为提高传统分类器在不平衡数据集上的分类性能,针对大多数过采样算法仅利用少数类数据的局部信息进行样本少数类生成,使得新生成的少数类样本不能很好地遵循原始少数类样本的分布,具有较少的信息量等问题,提出一种基于稀疏表示不平衡数据的过采样算法。首先利用少数类样本分布的全局信息进行样本合成,其次在样本确认阶段对合成的样本进行清洗,剔除位于多数类样本范围内的合成样本,防止噪声信息的传播。通过在 6 个不平衡数据集上与其他算法的性能比较,表明本文算法可以有效解决数据失衡问题,提高不平衡数据集的分类性能。

参 考 文 献

[1] Yang Q, Wu X. 10 challenging problems in data mining research[J]. International Journal of Information Technology and Decision Making, 2006, 5(4): 597 – 604.

[2] Ohsaki M, Wang P, Matsuda K, et al. Confusion matrix-based kernel logistic regression for imbalanced data classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(9): 1806 – 1819.

[3] Chawla N, Bowyer K W, Hall L O. SMOTE: synthetic minority over-sampling techniques[J]. Journal of Artificial Intelligence Research, 2002, 16: 321 – 357.

[4] Han H, Wang W Y, Mao B H, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing, 2005: 878 – 887.

[5] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-Level-SMOTE: Safe-level-syntheticminority over-sampling technique for handing the class imbalanced problem[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009: 475 – 482.

[6] Dong Y, Wang X. A new over-sampling approach: random-SMOTE for learning from imbalanced data sets[C]//International Conference on Knowledge Science, Engineering and Management, 2011: 342 – 352.

[7] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.

(下转第 306 页)

实验,结果表明,本文算法对于不同维度的数据集均实现理想的聚类准确率,并且对高维小样本数据实现较高的计算效率。但本文算法对于高维大样本数据的时间效率较为一般,未来将专注于本文算法应用于高维大样本数据集的研究,在提高聚类准确率的前提下,保持较快的处理速度。

参 考 文 献

[1] 宁可,孙同晶,徐洁洁. 面向海量数据的改进最近邻优先吸收聚类算法[J]. 计算机工程,2018,44(4):35 - 40.

[2] 杨洁,王国胤,庞紫玲. 密度峰值聚类相关问题的研究[J]. 南京大学学报(自然科学),2017,53(4):791 - 801.

[3] 贾声声,彭敦陆. CNN 支持下的领域文本自组织映射神经网络聚类算法[J]. 小型微型计算机系统,2018,39(6):1195 - 1200.

[4] Chang X, Wang Q, Liu Y, et al. Sparse regularization in fuzzy c-Means for High-Dimensional data clustering[J]. IEEE Transactions on Cybernetics,2017,47(9):2616 - 2627.

[5] 孙刚. 面向高维微阵列数据的集成特征选择算法[J]. 计算机工程与科学,2016,38(7):1330 - 1337.

[6] 李新玉,徐桂云,任世锦,等. 基于可靠性的正则化加权软k-均值的子空间聚类[J]. 南京大学学报(自然科学),2017,53(3):525 - 536.

[7] Tang Y, Browne R P, Mcnicholas P D. Model based clustering of high-dimensional binary data[J]. Computational Statistics & Data Analysis, 2015, 87(1):84 - 101.

[8] Huang J J. Using topic and subjectivity analysis for overlapped co-clustering documents[C]//2017 IEEE Third International Conference on Multimedia Big Data(BigMM), 2017.

[9] França F O D. A hash-based co-clustering algorithm for categorical data[J]. Expert Systems with Applications, 2016, 64(1):24 - 35.

[10] Yan Y, Chen L, Tjhi W C. Fuzzy semi-supervised co-clustering for text documents[J]. Fuzzy Sets & Systems, 2013, 215(215):74 - 89.

[11] Yao X, Han J, Zhang D, et al. Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering[J]. IEEE Transactions on Image Processing, 2017, 26(7):3196 - 3209.

[12] 肖辉辉,万常选,段艳明,等. 基于引力搜索机制的花朵授粉算法[J]. 自动化学报,2017,43(4):576 - 594.

[13] Bisson G, Hussain F. Chi-Sim: a new similarity measure for the co-clustering task[C]//2008 Seventh International Conference on Machine Learning and Applications. IEEE,2008.

[14] 孙娜,刘继文,肖东亮. 基于 BFGS 拟牛顿法的压缩感知 SL0 重构算法[J]. 电子与信息学报,2018,40(10):127 - 133.

[15] Arthur D, Vassilvitskii S. k-Means++: the advantages of careful seeding[J]. Proceedings of the Eighteenth Annual ACM-SIAM Symposiumon Discrete algorithms, Society for Industrial and Applied Mathematics, 2007,11(6):1027 - 1035.

[16] Singh A, Deep K. Hybridizing gravitational search algorithm with real coded genetic algorithms for structural engineering design problem[J]. Opsearch, 2017, 54(3):1 - 32.

[17] Mirjalili S, Gandomi A H. Chaotic gravitational constants for the gravitational search algorithm[J]. Applied Soft Computing, 2017, 53:407 - 419.

[18] Pelusi D, Mascella R, Tallini L, et al. Neural network and fuzzy system for the tuning of Gravitational Search Algorithm parameters[J]. Expert Systems with Applications, 2018, 102: 234 - 244.

[19] Bryant A, Cios K. RNN-DBSCAN: a density-based clustering algorithm using reverse nearest neighbor density estimates[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, 30(6):1109 - 1121.

[20] Wang Y, Liu X, Xiang L. GA-Based membrane evolutionary algorithm for ensemble clustering[J]. Comput Intell Neurosci, 2017, 2017(3):4367342.

[21] Farid D M, Nowe A, Manderick B. A feature grouping method for ensemble clustering of high-dimensional genomic big data[C]//Future Technologies Conference,2017.

(上接第 294 页)

[8] Nekooeimehr I, Lai-Yuen S K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets[J]. Ecpert Systems with Applications,2016,46:405 - 416.

[9] Yang A Y, Sastry S S, Ganesh A, et al. Fast L1-minimization algorithms and an application in robust face recognition: a review[C]//2010 IEEE International Conference on Image Processing,2010.

[10] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transaction on Systems, Man and Cybernetics, Part B(Cybernetics), 2009, 39(2):539 - 550.

[11] Starck J L, Elad M, Donoho D L, et al. Image decomposition via the combination of sparse representations and a variational approach[J]. IEEE Transactions on Image Processing, 2005,14(10):1570 - 1582.

[12] Tsaig Y, Donoho D L. Extensions of compressed sensing[J]. Signal Processing,2006,86(3):549 - 571.