一种基于非负矩阵分解的聚类集成算法

何梦娇 杨 燕 王淑营
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 为了解决通过原始数据集获得的基聚类结果存在一定的信息丢失,从而使得集成阶段的有效信息减少的问题,提出了一种基于非负矩阵分解的K-means聚类集成算法。该算法先利用K-means聚类算法获得集成信息矩阵,然后从原始数据集获取数据相关性,将两者结合后通过非负矩阵分解(NMF)技术构建共识函数以获得最终结果。实验证明,所提算法可以有效获取原始数据的潜在信息,并提高聚类质量。

关键词 聚类集成,K-means,NMF,潜在信息
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.011

NMF-Based Clustering Ensemble Algorithm

HE Meng-jiao YANG Yan WANG Shu-ying
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract A NMF-based K-means clustering ensemble (NBKCE) algorithm was proposed for solving the problem of effective information loss in ensemble, which is caused by basic clustering results obtained from the original datasets. In NBKCE, an ensemble information matrix is built primarily by exploiting the results of the K-means, and then the relationship matrix is formed based on the original dataset. At last nonnegative matrix factorization (NMF) is employed to construct consensus function to gain the final results. The experiments demonstrate that the NBKCE may attain the underlying information effectively and improve the performance of the clustering.

Keywords Ensemble clustering, K-means, NMF, Underlying information

聚类分析是数据挖掘的重要工具,现已被广泛应用于信息检索、医学图像处理、商务应用、Web搜索等领域。聚类,就是最大化一个簇内数据对象的相似度,最小化簇间的相似度。

虽然目前已经提出了很多聚类算法和相应的改进算法,但仍然存在较多问题。比如聚类结果在很大程度上还是取决于参数和初始化,且很难判断数据集中真实簇的个数,不同的聚类算法对同一个数据集进行聚类可能会产生不同的结果[1],即没有一种聚类算法能够准确地挖掘出各数据集所呈现的不同形状和结构的簇。因此,使用单一聚类算法进行聚类是一项比较困难的工作。

聚类集成将不同算法或同一个算法的多次运行结果进行合并,使得最终结果优于单个聚类算法的结果。相比于单一聚类算法,它在合并的过程中通常能够更好地发现数据集的许多结构和特征,可以解决单个聚类算法无法解决的问题,还能消除孤立点和噪声对聚类结果的影响。聚类集成能够达到任何一个单一的聚类算法都不可能达到的效果,因此研究聚类集成方法有着非常广泛的科研价值和实际应用价值。

Strehl等人[2]于2002年首次提出聚类集成概念,并基于图划分的思想提出了CSPA,HGPA,MCLA 3种集成算法,有效提高了聚类质量。文献[3]把每一个基聚类器看成是原数据的一个属性,并在此基础上提出了LVCE概率模型以及基于此模型的MCMC算法,它们都取得了良好的聚类效果。文献[4]通过建立投票机制来解决集成问题,同时根据投票结果得到聚类集成结果。基于多蚁群算法,文献[5]提出了一种集成方法,并取得了不错的聚类效果。而基于链接的聚类集成方法,文献[6]提出了簇间相关性的计算方法,从而获得更多的潜在信息,提高了聚类集成的质量。

本文受基于链接的模糊聚类集成方法的启发[7],提出了一种基于非负矩阵分解的K-means聚类集成算法(NMF-Based K-means Clustering Ensemble, NBKCE)。该算法将从原始数据集分别获取的相关性矩阵和信息矩阵相结合,利用NMF技术形成共识矩阵,迭代得到聚类集成结果,有效地获取了数据集的潜在信息,从而提高了聚类质量。

1 聚类原理

聚类分析是数据挖掘最重要的工具之一,在对数据对象的结构一无所知的情况下,利用数据对象本身的特性进行聚类,能够发现数据内在的结构,并探索其内在联系。

到稿日期:2016-07-16 返修日期:2016-08-30 本文受国家自然科学基金项目(61572407),国家科技支撑计划课题项目(2015BAH19F02)资助。

何梦娇(1991-),女,硕士生,主要研究方向为多视图聚类、聚类集成;**杨 燕**(1964-),女,博士,教授,博士生导师,主要研究方向为数据挖掘、计算智能、集成学习,E-mail:yyang@swjtu.edu.cn(通信作者);**王淑营**(1974-),女,博士,研究员,主要研究方向为云服务平台架构、自适应演化技术。

聚类的主要目标是根据数据集自身的特性将其划分到各个簇中，使得同一个簇中对象间的相似度最大化，而簇间的相似度尽可能小。根据聚类算法的主要思路，传统的聚类方法可分为 5 种[8]：划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。

1.1 非负矩阵分解

NMF 在分解后能够保留更多原来样本所反映的信息。分解后得到的结果是非负的，具有很好的物理意义，且实现过程简单快捷。顾名思义，NMF 就是将一个非负的矩阵分解成两个非负矩阵，并且这两个矩阵相乘的结果等于分解前的原矩阵。其目标函数如式(1)所示：

$$\begin{aligned}&\min \ \| X-WH \|_F^2 \\ &\text{s.t. } W,H\geqslant 0\end{aligned} \tag{1}$$

其中，非负的数据集 $X\in R^{m*n}$ 是原始矩阵，$X_{m*n}=(x_1,x_2,\cdots,x_n)$，$x_i$ 表示一个 m 维的列向量，即一个样本的信息。基矩阵 $W\in R^{m*r}$，$W_{m*r}=(w_1,w_2,\cdots,w_r)$，$w_i$ 表示一个 m 维的列向量，代表一个基向量。系数矩阵 $H\in R^{r*n}$，$H_{r*n}=(h_1,h_2,\cdots,h_n)$，其中 h_i 是 r 维的列向量，可以看作是 x_i 向量投影在 W 基矩阵定义的新空间中的坐标，满足 $x_i=W*h_i$，x_i 是投影系数。r 满足条件 $(m+n)*r<m*n$，即将一个高维的非负矩阵分解成两个低秩非负矩阵的乘积。迭代规则如式(2)、式(3)所示，其中⊙表示哈达码程式。

$$W\leftarrow W\odot\frac{(XH^{\mathrm{T}})_{ij}}{(WHH^{\mathrm{T}})_{ij}} \tag{2}$$

$$H\leftarrow H\odot\frac{(W^{\mathrm{T}}X)_{ij}}{(W^{\mathrm{T}}WX)_{ij}} \tag{3}$$

在 NMF 迭代过程中，基矩阵没有约束条件，数据之间存在着大量的冗余。为此，近年来提出了很多关于 NMF 的改进算法。

1.2 基于 NMF 的 K-means 聚类

Ding 等[9]将 K-means 聚类算法转换成了矩阵形式，如式(4)所示：

$$\begin{aligned}&\min_{F,G} \| X^{\mathrm{T}}-GF^{\mathrm{T}} \|_F^2 \\ &\text{s.t. } G_{ik}\in\{0,1\},\sum_{k=1}^{K}G_{ik}=1,\forall i=1,2,\cdots,n\end{aligned} \tag{4}$$

其中，$X=(x_1,x_2,\cdots,x_n)\in R^{d*n}$ 为原始数据集矩阵，拥有 n 个样本，每个样本有 d 维特征。$F\in R^{d*K}$ 为聚类中心矩阵，$G\in R^{n*K}$ 是隶属度矩阵，每行只有一个值为 1，其余皆为 0。即如果样本 x_i 属于第 k 个簇，则 $G_{ik}=1$。

1.3 聚类集成

聚类集成的过程主要包括两个阶段：1)假设有 n 个原始的数据对象集合 $X=(x_1,x_2,\cdots,x_n)$，对其运行 M 次初始化设置不同的同种聚类算法，或者采用几种普通的聚类算法得到 M 个有差异性的结果，表示为 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$；2)设计共识函数阶段，其目的是对 M 个有差异性的聚类结果 Π 进行融合，得到对于数据集合 X 的一个新的数据划分，集成结果一般会比单一聚类算法的结果好。

2 基于非负矩阵分解的 K-means 聚类集成方法

本文提出了一种基于非负矩阵分解的 K-means 聚类集成算法(NBKCE)，该算法将来源于原始数据集的关系矩阵与集成第一阶段产生的信息矩阵相结合，通过 NMF 技术，利用联系两者的隶属度矩阵进行聚类集成，有效地挖掘原始数据的潜在信息，从而提高聚类集成的质量。

2.1 NBKCE 算法

基于非负矩阵分解的 K-means 聚类集成算法(NBKCE)首先需要获取集成信息矩阵。假设有数据集 $X_{m*n}=(x_1,x_2,\cdots,x_n)$，用 K-means 聚类算法运行 M 次得到 M 个有差异性的聚类结果集 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$，其中 $\pi_i(i=1,2,\cdots,M)$ 表示第 i 个 K-means 基聚类结果。然后将聚类结果表示成一个 $n*M$ 的信息矩阵，即 $\Psi(M)=(\pi_1,\pi_2,\cdots,\pi_M)$。数据相关性矩阵用 $S\in R^{n*n}$ 表示，用来指导集成阶段的聚类，定义如式(5)所示：

$$S_{ij}=\frac{x_i(x_j)^{\mathrm{T}}}{\| x_i \| * \| x_j \|} \tag{5}$$

受 NMFCSJ[10]的启发，NBKCE 算法利用 NMF 对信息矩阵和关系矩阵进行分解，$\Psi(M)^{\mathrm{T}}=F\tilde{G}$，从而获得隶属度矩阵，同时受到关系矩阵 $S=G^{\mathrm{T}}G$ 的约束，利用更多原始数据集信息帮助完成聚类集成。其共识函数如下：

$$\begin{aligned}&\min_{F,\tilde{G},G} \| \Psi(M)^{\mathrm{T}}-F\tilde{G} \|_F^2+\lambda \| S-G^{\mathrm{T}}G \|_F^2 \\ &F\geqslant 0,\tilde{G}\geqslant 0,G\geqslant 0\end{aligned} \tag{6}$$

其中，$F\in R^{M*K}$ 为聚类中心矩阵，$\tilde{G}\in R^{K*n}$ 是隶属度矩阵，λ 是关系矩阵的权重。$G_{ij}=\tilde{G}_{ij}/\sum_{i=1}^{K}\tilde{G}_{ij}$[11]，然后将 G 中每列的最大值 G_{ij} 记为 1，同列其余值记为 0。该过程可以利用对角矩阵 U 实现，令 $U_{ii}=\sum_{j=1}^{K}\tilde{G}_{ij}$，则可以简单地转化为 $\tilde{G}=GU$。

综上，可以将共识函数转化为式(7)：

$$\begin{aligned}&J=\min_{F,\tilde{G},G} \| \Psi(M)^{\mathrm{T}}-FGU \|_F^2+\lambda \| S-G^{\mathrm{T}}G \|_F^2 \\ &\text{s.t. } G_{ij}\in\{0,1\},\sum_{i=1}^{K}G_{ij}=1,\forall j=1,2,\cdots,n \\ &u_{ij}\begin{cases}\geqslant 0, & i=j\\ =0, & \text{其他}\end{cases}\end{aligned} \tag{7}$$

通过对共识函数进行求解，迭代得到最终的隶属度矩阵。其中，$\Psi(M)\in R_+^{n*M}$ 为信息矩阵，$F\in R^{M*K}$ 为聚类中心矩阵，$G\in R^{K*n}$ 为隶属度矩阵，$S\in R^{n*n}$ 为数据相关性矩阵。

2.2 NBKCE 算法的优化过程

NBKCE 的实现过程实质上就是求解 G,F,U 的过程，只需要迭代求解这 3 个矩阵，直至目标函数收敛即可。因此，可以将矩阵迭代求解过程表示为如下 3 个步骤：

1)给定 G,U，求解 F；

2)给定 G,F，求解 U；

3)给定 F,U，求解 G。

求算矩阵 F，相当于最优化式(8)，比较式(8)与式(1)，可以简单类比推出 F 的迭代优化式(9)。

$$\min_{F,\tilde{G},G} \| \Psi(M)^{\mathrm{T}}-FGU \|_F^2 \tag{8}$$

$$F\leftarrow F\odot\frac{(\Psi(M)^{\mathrm{T}}U^{\mathrm{T}}FG^{\mathrm{T}})_{ij}}{(FGUU^{\mathrm{T}}G^{\mathrm{T}})_{ij}} \tag{9}$$

对于对角矩阵 U，可以简单套用最小二乘法直接得到迭代式(10)。

$$u_{ii}=\frac{(FG)_i^{\mathrm{T}}\Psi(M)_i}{(FG)_i^{\mathrm{T}}(FG)_i} \tag{10}$$

得到聚类中心矩阵 F 和对角矩阵 U 后，再对隶属度矩阵 G 的迭代公式进行推导。先将目标函数 J 对 G 求导，得到式(11)。

$$\frac{\partial J}{\partial G}=2F^{\mathrm{T}}FGUU^{\mathrm{T}}-2F^{\mathrm{T}}\Psi(M)U-2\lambda GS+2\lambda G(G^{\mathrm{T}}G) \tag{11}$$

根据乘性迭代原理，可以得到隶属度矩阵 G 的迭代式(12)。

$$G\leftarrow G\odot\left(\frac{F^{\mathrm{T}}\Psi(M)U+\lambda GS}{F^{\mathrm{T}}FGUU^{\mathrm{T}}+\lambda G(G^{\mathrm{T}}G)}\right)^{\frac{1}{2}} \tag{12}$$

NBKCE 算法的简要实现过程如算法 1 所示。

算法 1 NMF-Based K-means Clustering Ensemble

输入：数据集 X，簇个数 K，关系权重 λ，迭代次数 Iternum

输出：数据对象标签

1. 初始化 G，F，U
2. 对数据集 X 运行 M 个 K-means 聚类算法，得到 M 个不同的结果集 π_i，将其表示成信息矩阵 $\Psi(M)=(\pi_1,\pi_2,\cdots,\pi_M)$；
3. 根据式(5)获取关系矩阵 S；
4. For i=1：Iternum
5. 根据式(9)求解 F；
6. 根据式(10)求解 U；
7. 根据式(12)求解 G；
8. 得到目标函数值 J；
9. End

3 实验结果与分析

3.1 测试数据集选取

本次实验选取人工数据集和 UCI[12] 数据集两种来源的 10 个数据，其中 2d4c 是基于高斯分布随机产生的人工数据集，其余皆来源于 UCI 的真实数据集，其中 balance，heart，liver，cmc 分别是数据集 balance-scale，heart-statlog，liver disorders，contraceptive-method-choice 的缩写。所有测试数据集的相关统计信息如表 1 所列。

表 1 实验测试数据集的相关信息描述

数据集	样本个数	属性	分类数	来源
2d4c	200	2	4	人工
wine	178	13	3	UCI
iris	150	4	3	UCI
glass	214	9	6	UCI
segment	2310	19	7	UCI
balance	627	2	3	UCI
diabetes	416	9	2	UCI
heart	270	13	3	UCI
liver	345	6	2	UCI
cmc	1473	9	3	UCI

3.2 实验设计

设置基聚类器 K-means 的运行次数 $M=10$，并将得到的不同结果集组成信息矩阵来进行实验。将关系矩阵权重参数 λ 设为 0.0001，该值由经验得出。

将本文算法与 5 种集成算法进行实验比较，这 5 种算法分别为 CSPA，HGPA，MCLA[2]，HGBF[13] 和 EMcN[14]。

3.3 评价标准

本次实验将采用 F1[15] 和 RI(Rand index)[16] 对实验结果进行评价。

F1 的定义如下：

$$F1=\frac{2*PR}{P+R} \tag{13}$$

其中，P 为精确率，表示提取出的正确对象占提取出的对象的比例；R 为召回率，表示提取出的正确对象占样本的比例。

RI 的定义如下：

$$RI(\Pi,\pi)=\frac{n_{11}+n_{00}}{n_{11}+n_{10}+n_{01}+n_{00}} \tag{14}$$

其中，Π 为真实数据集，π 为聚类结果标签。n_{11} 表示数据对象在 Π 和 π 集合中都在一个簇中的个数，n_{01} 表示在 π 集合是在同一簇但是在 Π 的不同簇的个数，n_{00} 和 n_{10} 表示的含义同理。

根据上述定义可知，F1 和 RI 的值越大，聚类效果越好。

3.4 NBKCE 收敛性分析

NBKCE 算法基于乘性迭代原理，其收敛性由实验可以证明，如图 1 所示。根据式(7)，实验针对各个数据集求得相应的共识函数 J 的值。图 1 中 Jbalance 显示的是数据集 balance在不同迭代次数下的实验结果。

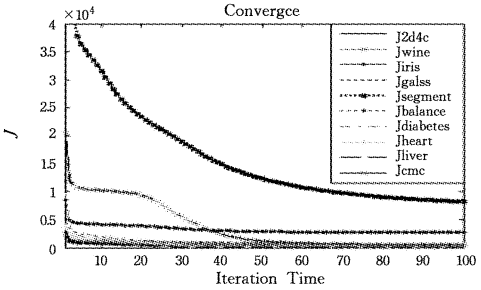

图 1 算法的收敛性

从图 1 可以看出，各个数据集的 J 值都随着迭代次数的增加而逐渐下降。大部分数据集的共识函数在迭代 10 次内迅速收敛，少部分的数据集在迭代 55 次时收敛，由此可以证明 NBKCE 是收敛的。

3.5 实验结果分析

表 2 与表 3 分别列出了运行 20 次后各种算法在每个数据集上的平均 F1 值和 RI 值，加粗的数值为同一评价指标下聚类结果最优的算法所产生的聚类结果。

表 2 各种算法的平均 F1 值

数据集	CSPA	HGPA	HBGF	EMcN	MCLA	NBKCE
2D4C	0.9450	0.2855	0.9560	**0.9700**	**0.9700**	**0.9700**
wine	0.6686	0.4034	0.7142	**0.7148**	**0.7148**	0.6884
iris	**0.9000**	0.4000	0.8880	0.8918	0.8918	0.8918
glass	0.4603	0.2445	0.5005	0.5178	0.4558	**0.5242**
segment	0.5610	0.1649	0.5414	0.5804	**0.5816**	**0.5816**
balance	0.5412	0.4914	0.5590	0.5670	0.5746	**0.6467**
diabetes	0.5492	0.5142	0.5765	0.6328	0.6328	**0.6428**
heart	0.4979	0.5024	0.4979	0.4905	0.4905	**0.5035**
liver	0.5680	0.5162	0.5657	**0.6471**	**0.6471**	**0.6471**
cmc	0.4007	0.3400	**0.5986**	0.4027	0.4020	0.4013
最优次数	1	0	1	3	4	7
最劣次数	0	9	0	1	1	0

表 3　各种算法的平均 RI 值

数据集	CSPA	HGPA	HGBF	EMcN	MCLA	NBKCE
2D4C	0.9478	0.6271	0.9581	**0.9708**	**0.9708**	**0.9708**
wine	0.6836	0.5379	**0.7247**	0.7187	0.7187	0.7194
iris	**0.8859**	0.5416	0.8846	0.8797	0.8797	0.8797
glass	0.7233	0.6481	0.7155	0.6975	0.7054	**0.7234**
segment	0.8328	0.7554	0.8362	0.8369	0.8296	**0.8373**
balance	0.5780	0.5339	0.5825	0.5845	0.5919	**0.6271**
diabetes	0.4886	0.4994	0.5082	**0.5507**	**0.5507**	**0.5507**
heart	0.5079	0.5014	0.5097	0.5041	0.5041	**0.5141**
liver	0.5071	0.4989	0.5012	0.5012	0.5012	**0.5043**
cmc	0.5575	0.5484	**0.6374**	0.5577	0.5580	0.5582
最优次数	1	0	2	2	2	7
最劣次数	1	9	0	1	1	0

在 F1 评价指标下，NBKCE 算法在 10 个数据集中取得了 7 次最优结果，0 次最劣结果，相较于对比算法有效地提高了聚类集成的质量；在 R2 评价指标下，其取得了 7 次最优结果，在其他 3 个数据集中也取得了较好的结果。由此可以看出，NBKCE 能够有效使用数据集的潜在信息来提高聚类质量。

结束语　本文提出了一种基于非负矩阵分解（NMF）的聚类集成方法 NBKCE。该算法将来自于原数据集的关系矩阵与信息矩阵结合后融入到共识函数中，利用 NMF 技术获取隶属度矩阵，有效利用潜在信息，提高了聚类集成的性能。今后的工作将考虑加入部分监督信息以改善集成效果，同时考虑将该聚类集成算法应用于多视图聚类集成的研究，以提高多视图聚类的性能。

参 考 文 献

[1] YANG C Y, LIU D Y, YANG B, et al. The research on clustering ensemble[J]. Computer Science, 2011, 38(2): 166-170. (in Chinees)
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.

[2] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[3] WANG H J, LI Z S, CHENG Y, et al. A Latent Variable Model for Cluster Ensemble[J]. Journal of Software, 2009, 20(4): 825-833. (in Chinees)
王红军，李志蜀，成飚，等. 基于隐含变量的聚类集成模型[J]. 软件学报，2009，20(4)：825-833.

[4] ZHOU Z H. Ensemble Methods: Foundations and Algorithms [M]. Taylor & Francis, 2012.

[5] YANG Y, KAMEL M. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern Recognition, 2006, 38(7): 1278-1289.

[6] LAMON N, BOONGOEN T, GARRETT S. Link-based cluster ensemble approach for categorical data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 413-425.

[7] YANG Y, FENG C F, JIA Z, et al. A Link-Based Fuzzy Clustering Ensemble[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(6): 887-892. (in Chinees)
杨燕，冯晨菲，贾真，等. 基于链接的模糊聚类集成方法[J]. 电子科技大学学报，2014，43(6)：887-892.

[8] HAN J, KAMBER M. Data Mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2006, 5(4): 1-18.

[9] DING C, HE X, SIMON H. Nonnegative lagrangian relaxation of k-means and spectral clustering[C]// ECML. 2005: 530-538.

[10] ZHANG J S, WANG C P, YANG Y Q. Learning latent features by nonnegative matrix factorization combining similarity judgments[J]. Neurocomputing, 2015, 155: 43-52.

[11] MIAO L D, QI H R. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization[J]. IEEE Trans. Geosci. Remote Sens., 2007, 45(3): 765-777.

[12] ASUNCION A, NEWMAN D J. UCI machine learning repository school of information and computer science, university of california [DB/OL]. (2007-06-02). http://www.ics.uci.edu/~mlearn/MLRepository.html.

[13] FERN X Z, BRODLEY C E. Solving cluster ensemble problems by bipartite graph partitioning[C]// Proc. 21th Int. Conf. Mach. Learn.. 2004: 36-44.

[14] ALEXANDER T, ANIL K J, WILLIAM P. Clustering Ensembles: Models of Consensus and Weak Partitions[J]. IEEETrans. on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[15] YANG Y, JIN F, KAMEL M. Survey of clustering validity evaluation[J]. Application Research of Computer, 2008, 25(6): 1630-1632. (in Chinese).
杨燕，靳蕃，KAMEL M. 聚类有效性评价综述[J]. 计算机应用研究，2008，25(6)：1630-1632.

[16] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of American Statistical Association, 1971, 66(336): 846-850.

（上接第 44 页）

[22] YAO Y Y. Interval sets and interval-set algebras[C]// IEEE International Conference on Cognitive Informatics (ICCI 2009). Hong Kong, China, 2009: 307-314.

[23] ZADEH L A. Fuzzy sets* [J]. Information & Control, 1965, 8(3): 338-353.

[24] PEDRYCZ W. Shadowed sets: representing and processing fuzzy sets[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 1998, 28(1): 103-109.

[25] ABDULLAH S, GOLAFSHAN L, ZAKREE M, et al. Re-heat simulated annealing algorithm for rough set attribute reduction [J]. International Journal of Physical Sciences, 2011(8): 2083-2089.

[26] CHEN Y P, EAMONN K, HU B, et al. The UCR Time Series Classification Archive[DB/OL]. http://www.cs.ucr.edu/~eamonn/time_series_data.