计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):273−278 [doi: 10.15888/j.cnki.csa.008184]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

多代表点的加权近邻分类算法①

林高思源

(福建师范大学 计算机与网络空间安全学院, 福州 350117)
通讯作者: 林高思源, E-mail: sylingao@gmail.com

摘　要: 传统的 KNN 算法存在分类效率低等缺点. 针对这些缺点, 本文提出一种高效的结合多代表点思想的加权 KNN 算法, 利用变精度粗糙集上下近似区域的概念, 结合聚类算法生成代表点集合构造分类模型, 再运用结构风险最小化理论优化分类模型并对影响分类模型的因素进行分析. 分类过程中根据测试样本与各代表点的相似度, 得到测试样本的相对位置. 其中属于样本点下近似区域的测试样本可直接判断其类别. 若测试样本在其他区域, 则根据测试样本与各代表点的相对位置对各代表点覆盖范围内的样本进行加权后判断测试样本的类别. 在文本分类领域的数据集上进行实验, 结果表明该算法能有效的提高分类模型的性能.

关键词: 近邻分类; 文本分类; 变精度粗糙集; 代表点; 分类模型; 样本加权

引用格式: 林高思源.多代表点的加权近邻分类算法.计算机系统应用,2021,30(12):273–278. http://www.c-s-a.org.cn/1003-3254/8184.html

Weighted Nearest Neighbor Classification Algorithm of Multi-Representative

LIN Gao-Si-Yuan

(College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)

Abstract: The traditional KNN algorithm has shortcomings such as low classification efficiency. This study proposes an efficient weighted KNN algorithm that combines the idea of multiple representative points. It uses the concept of the upper and lower approximate regions of the variable precision rough set and integrates the clustering algorithm to generate a representative point set and construct a classification model. Then it adopts the structural risk minimization theory to optimize the classification model and analyze the factors that affect the classification model. During the classification process, the relative position of the test sample is obtained according to the similarity between the test sample and each representative point. Moreover, the category of the test sample in the lower approximate region can be directly determined. If the test sample is in other areas, the sample within the coverage of each representative point is weighted according to the relative position of the test sample and each representative point to determine the type of the test sample. Experiments on the data set in the field of text classification show that the algorithm can improve the performance of the classification model.

Key words: nearest neighbor classification; text classification; variable precision rough set; representative; classification model; sample weighting

KNN 算法是一种基本的分类方法, 在文本分类、人脸识别[1]、遥感图象识别等各个领域都有着广泛的应用, 特别是当大规模数据集各类别分布区域重叠较多时, 较其他算法有更优良的表现.如戚玉娇等[2] 使用 KNN 算法对大兴安岭地区森林地上碳储量进行遥感估算; 宋飞扬等[3] 提出一种空气质量预测模型, 通过结

① 收稿时间: 2021-02-05; 修改时间: 2021-03-05; 采用时间: 2021-03-16

合 KNN 算法和 LSTM 模型, 有效提高模型的预测精度; 薛卫等[4] 使用 AdaBoost 算法对 KNN 算法分类器进行集成, 用于蛋白质的亚细胞定位预测问题.

虽然 KNN 算法简便易懂, 但亦存在一些缺点.为克服这些缺点, 目前已经有很多学者提出改进的方法, 如 Guo 等提出的 KNN Model 算法[5,6], 使用代表点集合来建立分类模型, 且能自动确定近邻数 k 的取值; 陈黎飞等[7] 提出一种多代表点的学习算法 MEC, 该算法基于结构化风险最小化理论[8] 来进行理论分析, 且使用聚类算法生成代表点集合, 以此构建分类模型. 刘继宇[9] 等提出一种基于普通粗糙集的加权 KNN 算法, 对现有的粗糙集值约简算法可能存在的问题进行改进. 王邦军等[10] 提出基于多协方差和李群结构的李-KNN 分类算法, 该算法利用集成的思想, 充分利用各种统计特征的影响, 结合 KNN 算法和李群结构的优势. 刘发升等[11] 提出一种基于变精度粗糙集的加权 KNN 文本分类算法 (VRSW-KNN 算法), 通过引入变精度粗糙集结合样本加权的方法来更有效的控制类别的分布区域, 从而有效提高分类的准确率, 然而比较依赖参数的取值.

根据以上研究现状, 本文提出一种基于多代表点的变精度粗糙集加权 KNN 算法, 并且使用该算法在文本分类领域进行应用. 本算法在 VRSW-KNN 算法的基础之上, 借鉴该算法的分类过程, 引入多代表点思想生成一系列代表点集合并使用聚类算法构建更加精确的分类模型, 使用结构化风险最小化理论进行理论层面的分析和进一步优化算法, 最终得出一个数目适宜的模型簇集合用于分类. 与 VRSW-KNN 算法相比, 本文算法不仅能更精确的划分各类别的分布区域, 更好的适应各种情况的数据集, 且能大幅降低分类的时间开销.

本文的组织结构如下: 第 1 节介绍本文的背景知识与相关工作; 第 2 节描述本文提出的算法, 并对部分影响因素进行理论层面分析; 第 3 节给出实验环境, 并对实验结果进行分析; 第 4 节总结全文, 同时给出未来可能的研究方向.

1 背景知识与相关工作

1.1 基于变精度粗糙集的加权 KNN 文本分类算法 (VRSW-KNN)

Ziarko[12] 于 1993 年利用相对错误分类率提出一种变精度粗糙集模型, 该模型利用精度 β 把原来经典粗糙集的上、下近似区域推广到任意精度水平, $\beta \in [0,0.5)$, 使粗糙集的区域变得更加灵活.

由于数据集的类别分布并不一定均衡, 这导致普通 KNN 算法在判断类别时会更容易把训练集中占比较多的类别给予测试样本, 使分类精确度降低. 为更好的解决该问题, 该算法根据测试样本的 K 近邻样本分布特征使用式 (1) 对 K 个样本分别计算权重, 再使用式 (2) 计算测试样本向量与 K 个训练样本向量的相似度, 最后使用式 (3) 计算各类别最终权重, 并将该向量分配给最终权重最大的类别[11].

样本权重计算公式[11]:

$$Weight_i = \frac{1}{\left(1+\dfrac{Num(C_i)}{Avgnum(C_l)}\right)^{\frac{1}{t}}},\ t>0 \tag{1}$$

式中,$Num(C_i)$为 K 近邻中属于C_i的文本数量; $Avgnum(C_l)$是指 K 近邻中类别的平均文本数.

相似度计算公式[11]:

$$sim\left(d_i,d_j\right) = \frac{\sum_{n=1}^{m} W_{in}\times W_{jn}}{\sqrt{\sum_{n=1}^{m} W_{in}\times W_{in}}\sqrt{\sum_{n=1}^{m} W_{jn}\times W_{jn}}} \tag{2}$$

其中, 特征向量$d_j=\{W_{i1},W_{i2},\cdots,W_{in}\}^{\mathrm{T}}$, 而$W_{in}$代表文件$d_j$的第$n$维.

VRSW-KNN 决策规则为[11]:

$$p(d_j,c_i) = \sum_{j=1}^{k} Weight_i \times sim(d_i,d_j)\times y(d_j,c_i)$$

$$y(d_j,c_i) = \begin{cases} 0, & d_j \notin c_i \\ 1, & d_j \in c_i \end{cases} \tag{3}$$

VRSW-KNN 算法利用变精度粗糙集的上下近似区域来使每个类别都有一个以类别质心作为中心点的近似区域, 以此检测测试样本和各类别分布区域的相对位置. 同时结合文本数量加权的思想, 从而能够在文本分类问题上比普通的 KNN 算法表现更好. 其计算类X_i的上下近似半径的算法步骤如下[11,13]:

Begin

Step 1. 计算第 i 类的质心 $O(X_i)$.

Step 2. 计算类别中心$O(X_i)$与同类别样本的相似度, 则第 i 类别的上近似半径为相似度的最低值.

Step 3. 计算类别中心$O(X_i)$与训练集中全部样本

的相似度, 如果样本相似度超过上近似半径, 则将其按照降序进入队列中.

Step 4. 依次将队列中样本移出. 直到$\frac{Num(n)}{n}>\beta$, 得下近似半径为第 n–1 个元素与中心$O(X_i)$的相似度.

End

算法中$Num(n)$表示前 n 个样本中不属于类别X_i的样本数目.

VRSW-KNN 算法总体步骤如下[11,13]:

Begin

Step 1. 数据预处理, 将文本数据转变为向量空间内的向量$d_i=\{W_{i1},W_{i2},\cdots,W_{in}\}^{\mathrm{T}}$.

Step 2. 得出各类别近似半径.

Step 3. 根据式 (2) 得出测试样本与各类别的相似度, 确定测试样本在数据集整体中的相对位置.

Step 4. 分类阶段应首先得出该样本是否属于某类别的下近似半径范围内, 如成立则为该样本添加这一类别的类别标记, 且算法终止.

Step 5. 如测试样本与一些类别中心的相似度大于类别的上近似半径, 则在这些类别中先使用式 (1) 计算各类别的权重, 再使用式 (3) 进行最终决策.

Step 6. 如测试样本与所有类别中心的相似度都小于类别的上近似半径, 则在完整训练集中先使用式 (1) 计算各类别的权重, 再使用式 (3) 来进行最终决策.

End

2 多代表点的 VRSW-KNN 算法

多代表点的 VRSW-KNN 算法, 全称 MVRSW-KNN, 简称 M-KNN, 该算法利用多代表点思想来让每个类别的代表区域变得更加细致, 从而通过增强分类模型的判断能力, 来解决 VRSW-KNN 算法在部分情况下表现失常的问题. 本节将先介绍模型簇的形式定义和总体分类模型, 根据结构风险理论[8] 来确定算法的优化目标, 再给出算法的总体过程并对此进行分析.

2.1 分类模型

M-KNN 算法构建分类模型是为能从给定数据集$Tr=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$中得到被优化过的$k$个类别的模型簇集合$\{C_1,C_2,\cdots,C_k\}$, 其中的$C_i$指的是第 i 类的模型簇集合, 可看作$C_i=\{p_{i1},p_{i2},\cdots,p_{im}\}$, 其中的$p_{ij}$是经过优化过的第 i 类模型簇集合的第 j 个模型簇, m 为第 i 类的模型簇总数, 同时可把模型簇p_{ij}覆盖区域内的样本点集合称为X_{ij}.

模型簇$p_{ij}=\left\{O\left(X_{ij}\right),\overline{R}_{ij},\underline{R}_{ij},Class_{ij}\right\}$, 其中$O\left(X_{ij}\right)$为模型簇$p_{ij}$的中心, $\overline{R}_{ij}$和$\underline{R}_{ij}$分别为模型簇p_{ij}的上近似半径和下近似半径, $Class_{ij}$为模型簇p_{ij}的类别标号.

模型簇p_{ij}的代表点$O\left(X_{ij}\right)$是该模型簇范围内所有点的中心, 因此亦可称为中心点, 采用式 (4) 计算:

$$O(X_{ij})=\frac{1}{|X_{ij}|}\sum_{x\in X_{ij}}x \tag{4}$$

其中, $|X_{ij}|$表示X_{ij}集合内包含的样本点数目. 此外模型簇p_{ij}的$\overline{R}_{ij},\underline{R}_{ij}$的计算方法与 VRSW-KNN 算法中上下近似半径的计算方法一致.

分类算法步骤如下:

输入: 模型簇集合$\{C_1,C_2,\cdots,C_k\}$, 待分类样本x_α, 参数 t, 近邻数 k.

输出: 样本数据x_α的类别y_α.

Begin

Step 1. 根据式 (2) 计算x_α与各模型簇的中心点之间的相似度, 从而确定x_α的位置.

Step 2. 首先判断x_α是否属于某个模型簇的下近似区域, 如果属于, 则直接将x_α归于该模型簇所属的类别, 算法终止.

Step 3. 如样本属于某些模型簇的上近似区域, 则在这些模型簇的覆盖区域中选出与x_α相似度最大的k个样本后使用式 (3) 得出x_α的类别.

Step 4. 如果样本不属于所有类别的上近似区域, 则在所有模型簇的覆盖区域中选出与x_α相似度最大的k个样本后使用式 (3) 得出x_α的类别.

End

2.2 结构风险

应用分类算法来对未分类样本进行分类可看作一个 k-分类模型[7] 的变型, 即:

$$M_k=\{p_{kl}|l=1,2,\cdots,\alpha \text{ and } Class_{kl}=k\}$$

和基于该分类模型的二类分类判断, 即:

$$Prediction(x_\alpha,M_k)=\begin{cases}k, & \text{if } \exists p_{kl}\in M_k\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

其中, x_α为待分类样本.

为提高模型的性能, 基于该文献所提到的 k-分类模型来对本节的分类模型基于 VC 维理论[8] 进行结构风险的分析, 可得应用本节算法的分类模型来进行分

类亦是一个二类分类过程, 而二类分类模型期望风险的上界[8], 期望风险$R(M_k)$可表示为:

$$R(M_k) \le R_{\text{emp}}(M_k) + VC\_confidence(h_k),$$

其中, h_k 表示M_k 的 VC 维; $VC_confidence(h_k)$表示 VC 置信度; $R_{\text{emp}}(M_k)$表示经验风险, 为分类模型使用训练集时的平均误差. 而在本节算法中, M_k的经验风险为:

$$R_{\text{emp}}(M_k) = \frac{1}{n}\left(\sum_{(x,y)\in Tr, y=k} I(k \ne Prediction(x_\alpha, M_k)) + \sum_{(x,y)\in Tr, y\ne k} I(k = Prediction(x_\alpha, M_k)) \right) \quad (6)$$

由此可得, 分类的性能与 VC 置信度和经验风险都密切相关. 其中 VC 置信度随$|M_k|$即模型簇数目的增加而出现单调递增的状态. 这点可由研究学习矢量量化 (Learning Vector Quantization, LVQ) 算法的 VC 维的文献 [14] 证明. 根据其定理 1 的结论可得 LVQ 的 VC 维是其“原型”数目的一个单调递增函数, 而本文算法所定义的模型簇可在文献 [7] 的基础上亦可看作一种 LVQ“原型”的扩展. 因此根据 VC 置信度的定义[8], VC 置信度是 VC 维的递增函数, 可得本文分类模型的 VC 置信度亦随$|M_k|$的增加而单调递增.

经验风险亦与$|M_k|$有关, 这点可从极端情况上观察得出, 当$|M_k|$与类别数目相等时, 一个类别仅有一个簇, 这种情况下, 多代表点思想就失去作用, 变为 VRSW-KNN 算法, 经验风险在该情况下具有最大值; 而当$|M_k|$与样本点的数目相等的时候, 每个样本点都构成一个模型簇, 这种情况下, 经验风险降低到 0. 结合这两种极端情况, 可以得出结论, 经验风险总体上随着模型簇数目的增加而呈减少趋势. 但并不表明减少的过程是单调递减, 这点从文献 [7] 中可以得到证明.

综上可得, $R(M_k)$的上限由 VC 置信度和经验风险共同决定, 而这两个因素又都与$|M_k|$值有关, 因此模型训练算法的目的便是得出一个适合的$|M_k|$值来尽量使得分类模型的期望风险达到较低值.

模型训练算法:

输入: 训练集Tr, 参数 β, 分类模型类别标号$k(k = 1,2,\cdots,K)$.

输出: 分类模型M_k.

Begin

Step 1. 构造初始分类模型$M_k = \{p_{k1}\}$, 设 X_0={Tr 中类别标号为 k 的样本}, 根据类别标号为 k 的样本计算出初始模型簇的上下近似区域和中心点, 从而得出p_{k1}的各个值.

Step 2. 计算初始模型的$R_{\text{emp}}(M_k)$.

Step 3. 如果$R_{\text{emp}}(M_k) = 0$或者$|M_k|$等于数据点的个数, 返回M_k, 算法终止.

Step 4. 使用 K-means 聚类算法[15] 对X_l中的样本进行聚类, 划分成$|M_k|+1$个集合$X_1, X_2, \cdots, X_{|M_k|+1}$.

Step 5. 构造新分类模型$M_k' = \{p_{ki} | i = 1,2,\cdots, |M_k|+1\}$并构造其中的新模型簇$p_{ki} = \left\{O(X_{ki}), \overline{R}_{ki}, \underline{R}_{ki}, k\right\}$.

Step 6. 使用式 (6) 计算M_k'的经验风险$R_{\text{emp}}(M_k')$. 如果$R_{\text{emp}}(M_k') \ge R_{\text{emp}}(M_k)$, 返回$M_k$, 算法结束. 否则, $M_k = M_k'$, 重复 Step 3 至 Step 5.

End

对于每个类别, 训练算法的过程是从 1 开始, 逐个递增模型簇数目, 直到经验风险$R_{\text{emp}}(M_k)$达到一个局部极小值或为 0 后, 停止递增并记录目前的该类别模型簇状况. 根据文献 [7] 提供的策略, 这里的局部极小值指的是第一个经验风险极小值, 这样可以在降低训练阶段耗时的同时, 使 VC 置信度和经验风险达到一种平衡. 本节算法亦沿用该策略. 最终, 对于每个类别, 都有一至多个模型簇表示该类别的分布情况, 从而比 VRSW-KNN 算法更擅于对样本进行精准分类.

3 实验与分析

实验选择 KNN、VRSW-KNN, 这两种基于最近邻思想的分类器作为比对的对象. 实验设备为: CPU 为 CORE i7-8750H2.20 GHz, 16 GB RAM, Windows 10 操作系统的计算机.

3.1 实验数据

实验数据使用复旦大学原计算机信息与技术系国际数据库中心自然语言处理小组的中文语料库[11] 中的 2400 篇作为数据集, 包括艺术, 历史, 计算机, 环境, 体育, 农业等共 8 个类别.

其中的 1600 篇作为训练样本, 800 篇作为测试样本. 在正式实验前, 对数据集进行分词, 去停用词, 特征提取后再使用 TF-IDF 特征词加权方式将所有文本变为特征向量.

3.2 实验结果与分析

实验的分类性能评价指标采用 Micro-$F1$ 指标来衡量分类器的分类精度, 同时对 3 种算法所使用的训

练集和测试集均一致, 且实验结果为多次运行后的平均值, 以避免随机因素导致的影响.

本实验中表 1 为固定 t=1.0, k=10, 粗糙集精度 β 的取值分别为 0.05、0.10、0.15 时 3 种算法的 $F1$ 值, 由于参数 t 与耗时无关, 因此表 2 为仅 β 取不同值时 3 种算法的耗时对比.

表 1 β 取不同值时 3 种算法的 Micro-$F1$ 值对比

类别	KNN算法	VRSW-KNN			M-KNN		
		β=0.05	β=0.1	β=0.15	β=0.05	β=0.1	β=0.15
艺术	0.98	0.99	0.99	0.99	0.97	0.97	0.97
历史	0.86	0.94	0.94	0.94	0.76	0.72	0.76
计算机	0.92	0.77	0.42	0.34	0.79	0.86	0.86
环境	0.99	0.93	0.20	0.14	0.97	0.94	0.94
农业	0.74	0.74	0.73	0.69	0.80	0.83	0.83
经济	0.94	0.83	0.83	0.79	0.86	0.79	0.80
政治	0.85	0.89	0.88	0.83	0.85	0.85	0.85
体育	0.70	0.63	0.63	0.63	0.84	0.85	0.85
平均	0.87	0.84	0.70	0.67	0.86	0.85	0.86

表 2 β 取不同值时 3 种算法的耗时对比

耗时	KNN算法	VRSW-KNN			M-KNN		
		β=0.05	β=0.1	β=0.15	β=0.05	β=0.1	β=0.15
总耗时(s)	8226.69	4539.53	3197.88	2904.07	1851.05	1410.58	1417.36
平均耗时(s)	10.28	5.67	3.99	3.63	2.31	1.76	1.77

由表 1 可知, 当粗糙集精度 β 从 0.05 逐渐增加到 0.15 时, 对于 VRSW-KNN 算法, 模型簇的下近似区域覆盖范围逐渐变大, 导致在分类时模型簇的 $F1$ 值逐渐降低. 但对于 M-KNN 算法, 虽下近似区域覆盖范围发生相同变化, 但由于在训练模型时, 每个类别分布区域由多个模型簇组成, 导致虽粗糙集精度 β 发生较大变化, 但对各类别的覆盖区域影响较小, 不会导致 $F1$ 值的大幅降低. 因此可知使用多代表点思想后, 生成的模型簇的表现与 VRSW-KNN 算法相比不仅准确率略有上升, 且在参数的选取上更加容易接近较优良的结果.

由表 2 可知, 普通的 KNN 算法耗时最长, M-KNN 算法在耗时上较 VRSW-KNN 算法大幅缩短. 原因是虽然 VRSW-KNN 算法使用变精度粗糙集的思想, 然而只简单的设定一个类别只有一个簇, 导致并没有充分的利用该分类模型, 其分类时间随着样本数的增加而急剧增长. 很多数据点仍需要多个类别甚至全部类别的模型簇内样本寻找 K 近邻, 但 M-KNN 算法结合多代表点的思想, 将每个类别分布区域划分成多个模型簇, 虽然导致训练分类模型的时长大幅增加, 但在分类阶段则获得各类别更精确的分布区域, 让很多数据点能够在分类算法的前两步就获得较准确的标记, 从而大幅降低数据分类的耗时, 导致在本节算法中, 大部分耗时是训练分类模型的耗时. 因此, 在更大规模的文本数据集中, 本节算法的分类性能会远好于 VRSW-KNN 算法和普通的 KNN 算法.

综上所述, 本文算法能更加有效的提升分类的效率和准确率.

4 结束语

本文提出一种较高效的基于多代表点思想的变精度粗糙集加权 KNN 算法, 并使用该算法在文本分类领域中进行实验. 实验结果不仅成功验证期望风险的理论分析的正确性, 更表明该算法的实际可行性, 进一步解决 VRSW-KNN 算法在刻画各个类别分布情况的不足, 从而在大幅降低时间开销的同时提高分类的准确率. 下一步的工作重点是围绕训练模型的各个影响因素, 探索在保证分类准确率且能降低训练期间时间开销的相关方法.

参考文献

1 毋雪雁, 王水花, 张煜东. K 最近邻算法理论与应用综述. 计算机工程与应用, 2017, 53(21): 1–7. [doi: 10.3778/j.issn.1002-8331.1707-0202]

2 戚玉娇, 李凤日. 基于 KNN 方法的大兴安岭地区森林地上

碳储量遥感估算. 林业科学, 2015, 51(5): 46–55.
3 宋飞扬, 铁治欣, 黄泽华, 等. 基于 KNN-LSTM 的 $PM_{2.5}$ 浓度预测模型. 计算机系统应用, 2020, 29(7): 193–198. [doi: 10.15888/j.cnki.csa.007490]
4 薛卫, 王雄飞, 赵南, 等. 集成改进 KNN 算法预测蛋白质亚细胞定位. 生物工程学报, 2017, 33(4): 683–691.
5 Guo GD, Wang H, Bell D, *et al*. KNN model-based approach in classification. OTM Confederated International Conferences on the Move to Meaningful Internet Systems. Catania: Springer, 2003. 986–996.
6 Guo GD, Wang H, Bell D, *et al*. Using *k*NN model for automatic text categorization. Soft Computing, 2006, 10(5): 423–430. [doi: 10.1007/s00500-005-0503-y]
7 陈黎飞, 郭躬德. 最近邻分类的多代表点学习算法. 模式识别与人工智能, 2011, 24(6): 882–888. [doi: 10.3969/j.issn.1003-6059.2011.06.023]
8 Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121–167. [doi: 10.1023/A:1009715923555]
9 刘继宇, 王强, 罗朝晖, 等. 基于粗糙集的加权 KNN 数据分类算法. 计算机科学, 2015, 42(10): 281–286.
10 王邦军, 李凡长, 张莉, 等. 基于改进协方差特征的李-KNN 分类算法. 模式识别与人工智能, 2014, 27(2): 173–178. [doi: 10.3969/j.issn.1003-6059.2014.02.012]
11 刘发升, 董清龙, 李文静. 变精度粗糙集的加权 KNN 文本分类算法. 计算机工程与设计, 2019, 40(5): 1339–1342, 1364.
12 Ziarko W. Variable precision rough set model. Journal of Computer and System Sciences, 1993, 46(1): 39–59. [doi: 10.1016/0022-0000(93)90048-2]
13 董清龙. 果蝇优化算法改进及其扩展研究 [硕士学位论文]. 赣州: 江西理工大学, 2019.
14 Crammer K, Gilad-Bachrach R, Navot A, *et al*. Margin analysis of the LVQ algorithm. Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2002. 479–486.
15 Kotsiantis S, Pintelas P. Recent advances in clustering: A brief survey. WSEAS Transactions on Information Science and Applications, 2004, 1(1): 73–81.