计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000102

基于量子进化算法的非平衡数据混合采样算法

杨 浩[1] 陈红梅[2]
1 西南交通大学云计算与智能技术高校重点实验室 成都 611756
2 西南交通大学信息科学与技术学院 成都 611756
(apologise@my.swjtu.edu.cn)

摘 要 欠采样和过采样是解决非平衡数据分类问题的常用方法。针对目前解决数据非平衡分布主要采用单一的采样方法可能会导致过拟合或重要样本丢失的问题,提出了一种基于量子进化算法的混合采样方法 MSQEA(Mixed-Sampling method based on Quantum Evolutionary Algorithm)。该方法对多数类和少数类样本分别进行编码,组成量子进化算法中的个体种群,然后通过迭代得到合适的候选采样子集。针对得到的候选采样子集,首先使用欠采样移除多数类样本,避免了后续的过采样方法合成过多冗余的少数类样本的问题,然后采用过采样方法对少数类样本进行过采样,得到一个平衡数据集。同时,为了有效地评价量子个体的适应度,使用聚类算法对原始数据集进行聚类,构建一个有效的验证集来评价个体。为了验证 MSQEA 算法的性能,在 KEEL 网站下载的非平衡数据集上,采用 SMO,J48 和 NB 等作为分类算法测试不同采样算法处理后的分类性能。实验结果表明,MSQEA 算法相比当前较为优秀的采样算法在多种分类器上具有更好的分类性能。

关键词:非平衡数据;量子进化算法;混合采样;分类
中图法分类号 TP391

Mixed-sampling Method for Imbalanced Data Based on Quantum Evolutionary Algorithm

YANG Hao[1] and CHEN HONG-mei[2]
1 Key Laboratory of Cloud Computing and Intelligent Technology,Southwest Jiaotong University,Chengdu 611756,China
2 School of Information Science and Technology,Southwest Jiaotong University,Chengdu 611756,China

Abstract The under-sampling and over-sampling are the common methods for solving the classification problem in an imbalanced data. This paper focuses on the overfitting or lose valuable samples problems brought by using a single sampling method. A mixed sampling method,namely MSQEA,based on quantum evolutionary algorithm is proposed. In MSQEA,the majority class samples and minority class samples are firstly encoded separately to form individuals of population in the quantum evolutionary algorithm,and then an appropriate candidate sampling subset is obtained through optimization iterations. After that,the majority samples in candidate subset are removed by under-sampling to avoid the problem of subsequent oversampling method to generate overmuch redundant samples. Then,an oversampling method is used to generate the minority samples. Additionally,in order to effectively evaluate the fitness of quantum individuals,clustering technique is used to cluster the dataset and the effective validation sets for the evaluation of individuals are obtained. Experiments are conducted to evaluate the performance of algorithm MSQEA. The imbalanced data sets are downloaded from KEEL website,and SMO,J48 and NB are used as classifiers to verify the performance of a classifier after data preprocessing by different sampling methods. Experimental results show that the classification performance of MSQEA is better than some state-of-the art sampling methods.

Keywords Imbalanced data,Quantum evolutionary algorithm,Mixed-sampling,Classification

1 引言

在大数据时代,数据挖掘与分析在各种决策领域中扮演着越来越重要的角色。而在各种数据挖掘技术中,分类是商业和工程问题中应用最广泛的技术之一,例如文本挖掘[1]、医疗诊断[2]、自然灾害预测[3]和欺诈检测[4]等。受现实因素的

到稿日期:2019-10-16 返修日期:2020-03-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61572406,61976182);四川省国际科技创新合作重点项目(2019YFH0097)
This work was supported by the National Natural Science Foundation of China (61572406,61976182) and Key Program for International S&T Cooperation of Sichuan Province (2019YFH0097).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

影响，在上述应用中所得到的数据集往往是非平衡的。非平衡数据集是指不同类别之间样本数量差异明显，导致类别之间样本分布不平衡。在二分类数据集中，数量多的类别称为多数类，数量较少的类别称为少数类。在分类任务中，分类器通过减小分类误差来最大化分类精确度。然而，传统的分类算法只适用于类别分布较为均衡的数据集，当面对非平衡数据集时，由于多数类样本占据绝大部分比例，分类器为了最大化全局分类精确度，容易将少数类样本错分为多数类，忽视了少数类样本的分类正确性。而在实际应用中，少数类样本却往往是我们更为关注的对象。其次，在非平衡数据集中往往也伴随着小析取项(Small disjuncts)[5]和类别重叠(Class overlapping)[6]等问题。这些问题也会影响分类器的分类性能。因此，传统的分类算法并不适用于非平衡数据集，如何在不降低分类器的全局分类精确度的前提下提高少数类样本的分类精度是学者研究的重要任务之一。

解决非平衡数据问题的方法主要分为基于算法层面和数据层面。由于数据集的多样性和复杂性，基于数据层面的方法是应用最广泛、最简单的方法之一。数据层面的方法主要利用采样方法对数据集进行处理，从而使得类别之间达到平衡。采样方法主要可以分为欠采样和过采样，过采样方法的本质是采取合适的策略增加少数类样本的数量，由于在过采样过程中会合成冗余的样本或错误的样本，往往会导致分类模型过拟合。为了改善合成样本的质量，学者们提出了一系列有效的过采样方法。Chawla 等提出了利用样本的近邻构造少数类样本的 SMOTE(Synthetic Minority Over-sampling TEchnique)算法[7]。该算法通过在少数类样本与其对应的近邻之间进行线性插值，来合成新的少数类样本。相比随机采样方法，SMOTE 通过合成新的样本，改善了少数类样本的分布，缓解了过拟合问题。但是 SMOTE 没有考虑类别的分布，针对所有的少数类样本进行采样容易合成错误的样本[8]。学者们提出了一些基于 SMOTE 的改进算法。Barua 等结合 SMOTE 和聚类算法提出了一种基于样本权重的过采样算法，对少数类样本进行聚类，从而得到若干个簇，然后对每个簇中样本权重较高的样本进行采样[6]。Zhu 等基于 K 近邻算法提出了一种针对多类别非平衡数据的采样算法 SMOM，该算法通过在采样时赋予每个采样点的 K 近邻不同的选择权重来合成样本，从而提高了合成样本的质量[8]。为了解决非平衡数据中出现的类别重叠问题和小析取项问题，Yang 等提出了一种基于样本局部密度的非平衡数据集成分类算法 LADBMOTE(combined the Local Area Density and Bagging for Minority Oversampling TEchnique)[9]，该方法通过计算样本的局部密度对样本进行采样，然后集成采样后的数据集，较大程度地提高了非平衡数据的分类性能。

欠采样方法的做法与过采样相反，通过减少多数类样本的数量来达到类别平衡。由于欠采样随机移除一部分多数类样本，因此不仅会导致部分价值高的多数类样本被移除，还会破坏原有多数类样本的分布。为了能够最大程度地保留具有代表性的多数类样本，同时使多数类与少数类保持平衡，Lin 等提出了一种基于聚类的欠采样方法 CBS(Clustering-Based Sampling)[5]，该方法通过设置聚类中心数等于少数类样本数来对多数类进行聚类，然后选取聚类中心样本所组成的集合作为新的多数类集合，有效避免了类别重叠和小析取项问题给采样带来的负面影响。Tsai 等提出了一种将聚类方法与样本选择技术[10-12]结合的欠采样方法[13]。该方法首先对多数类进行聚类分析，然后使用样本选择技术对不同簇中的样本进行选择，得到新的多数类样本集。针对随机欠采样方法随机移除多数类样本会丢失重要样本的问题，Shao 等利用量子进化算法对多数类进行欠采样，并在采样的过程中将预测精度作为适应度函数，在移除冗余多数类样本的同时达到最优的分类精度[14]。Li 等基于多目标粒子群提出了 AMSCO(Adaptive Multi-objective Swarm Crossover Optimization)算法[15]，该算法使用粒子群算法和 SMOTE 算法对少数类样本进行过采样，生成多个子集，然后对多个子集进行融合筛选，最后在迭代的过程中得到最优的采样倍率和 SMOTE 算法中最优的近邻数 K 值。

传统的采样算法在解决二分类问题时都只采用了单一的采样方法。文献[16]表明，单一的采样方法容易加剧过拟合或者造成重要样本丢失的问题。欠采样移除多数类中的样本，使得多数类样本与少数类样本数量达到相对平衡。由于少数类样本的数量非常少，因此需要移除大量的多数类样本，不恰当的欠采样方法会导致多数类丢失有潜在价值的样本，使得多数类的分类精度下降。同样，过采样会合成大量的冗余样本或者噪音样本，使得分类模型变得过拟合或效果不佳。其次，在基于 SMOTE 的过采样方法及其改进方法中，采样的对象为少数类中所有的样本。文献[17-18]表明，处于不同区域的样本，采样合成的样本质量也会不同，因此单一的采样方法往往并不能取得很好的分类性能。

针对上述问题，本文提出了一种基于量子进化算法的混合采样方法。基于二分类非平衡数据问题，由于过采样方法 LADBMOTE 能较好地处理小析取项和类别重叠问题，因此本文将 LADBMOTE 与欠采样方法相结合以对数据集进行混合采样，从而有效地避免单一采样方法带来的问题。首先，考虑数据集中的样本分布差异对采样效果的影响，利用量子进化算法找到合适的候选采样子集，然后对候选采样子集进行混合采样。其次，利用聚类算法对原始数据集进行聚类，构建一个适合的验证集。通过构建的验证集，利用分类器对量子个体进行评价，得到个体的适应度，使得量子个体能够根据适应度在迭代中进行调整，提高了找到最优解的可能性。

2　相关工作

2.1　LADBMOTE 基本原理

当数据集出现类别重叠和小析取项问题时，基于 SMOTE 的过采样方法及其改进方法在使用欧氏距离选取近邻进行采样时会降低采样的质量和效果。基于样本局部密度的非平衡数据分类方法 LADBMOTE 能够较好地处理类别重叠问题[5]和小析取项问题[6]。该方法提出了一种基于样本

局部密度的 K 近邻选择策略，选择采样点的 K 个近邻分别进行采样，形成多个平衡数据集，使用集成学习对采样后的多个平衡数据集进行分类处理。

在非平衡数据集 D 中，少数类样本集合为 D_1，多数类样本集合为 D_2。假定采样点为 x_1，S_i 为根据 LADBMOTE 中定义的 K 近邻计算策略所得到的 x_i 的 K 个少数类样本近邻集合。LADBMOTE 计算采样点 x_i 的 K 个近邻的具体步骤如下。

步骤 1　针对采样点 $x_i \in D_1$，计算所有的少数类样本 $\{x_j \mid x_j \neq x_i\}$ 与 x_i 之间的欧氏距离 d_{ij}。

步骤 2　求得 x_i 与 x_j 的中点 $\hat{x}_{ij}$，计算以 $\hat{x}_{ij}$ 为圆心、以 d_{ij} 为直径的圆内多数类样本点的个数 σ_{ij}，计算公式如式(1)所示：

$$\sigma_{ij}=\sum_{x_k \in D_2} \chi(dist(x_k,\hat{x}_{ij})-\frac{d_{ij}}{2}) \tag{1}$$

其中，$dist(x_k,\hat{x}_{ij})$ 为 x_k 到 $\hat{x}_{ij}$ 的距离。

$$\chi(x)=\begin{cases}1, & \text{if } x<0\\ 0, & \text{if } x\geqslant 0\end{cases} \tag{2}$$

步骤 3　根据步骤 1 得到的距离选取 K 个距离采样点 x_i 最近的少数类样本点 $\{x_{i1},x_{i2},\cdots,x_{iK}\}$，然后根据步骤 2 计算得到 $\{\sigma_{i1},\sigma_{i2},\cdots,\sigma_{iK}\}$，将其中 $\sigma=0$ 所对应的样本点加入到 S_i 中。如果 $\{\sigma_{i1},\sigma_{i2},\cdots,\sigma_{iK}\}$ 均为 0，则将 $\{x_{i1},x_{i2},\cdots,x_{iK}\}$ 中所有的样本加入到近邻集合 S_i 中，计算加入到 S_i 中的少数类样本个数 k_1，如果 $k_1=K$，则结束近邻计算策略。

步骤 4　求所有候选近邻 $\{x_j \mid x_j \in D_1, \sigma_{ij} \neq 0\}$ 的局部密度 ρ_{ij}。局部密度的计算方式如式(3)所示：

$$\rho_{ij}=\frac{\sigma_{ij}}{area(\hat{x}_{ij},d_{ij})} \tag{3}$$

其中，$area(\hat{x}_{ij},d_{ij})$ 是以 $\hat{x}_{ij}$ 为圆心、d_{ij} 为直径的圆的面积。

步骤 5　对所有 $\sigma_{ij}\neq 0$ 对应的样本点的局部密度进行升序排序，选择 $K-k_1$ 个局部密度 ρ 最小的少数类样本点，并将其加入到 S_i 中，结束近邻计算策略。

步骤 6　根据选取的近邻，对采样点进行采样。在对所有少数类样本进行采样后，得到多个平衡的数据集并使用分类器进行训练。最后，利用集成学习将多个分类器进行集成，使用测试集评估集成分类器的分类性能。

2.2　量子进化算法的基本原理

量子进化算法(Quantum Evolutionary Algorithm，QEA)[19]是由 Han 等提出的一种智能进化算法，该算法将量子理论和进化算法相结合，使用量子比特的概率幅来表示每个个体，从而使得个体可以表示为多种状态的叠加，这种特点使得该算法能够很好地解决复杂的组合优化问题。

在量子进化算法中，每个个体由多个量子比特位构成，每个量子比特位使用概率幅来表示 0 和 1 的叠加态，叠加态表示该量子比特位既可能表示为 0(记为 $|0\rangle$)，也可能表示为 1(记为 $|1\rangle$)，或者其中的中间态。一个量子比特位的表示如式(4)所示：

$$|\psi\rangle=\alpha|0\rangle+\beta|1\rangle \tag{4}$$

其中，α^2 为该量子比特位取 0 的概率，β^2 为量子比特位取 1 的概率，并且满足条件 $\alpha^2+\beta^2=1$。在观测这个量子比特位之前，它既可以表示为 1，也可以表示为 0。观测后，量子比特位便坍塌为稳定态，只能表示 0 或 1。因此，如果个体包含 n 个量子比特位，则一个个体可以同时存储和表示 2^n 种包含 0 和 1 的序列。

在量子进化算法中，假定个体的长度设置为 n，种群个体数量为 m，第 t 代种群的量子群体记为 $Q(t)=\{q_1^t,q_2^t,\cdots,q_k^t,\cdots,q_m^t\}$。其中，个体 k 表示为：

$$q_k^t=\begin{bmatrix}\alpha_1^t & \alpha_2^t & \cdots & \alpha_n^t\\ \beta_1^t & \beta_2^t & \cdots & \beta_n^t\end{bmatrix} \tag{5}$$

为了使个体能够朝着最优个体进化，量子进化算法采用量子旋转门对个体的量子比特位进行调整。最常用的量子门如下：

$$\begin{bmatrix}\alpha_i^{t+1}\\ \beta_i^{t+1}\end{bmatrix}=\begin{bmatrix}\cos(\theta_i) & -\sin(\theta_i)\\ \sin(\theta_i) & \cos(\theta_i)\end{bmatrix}\begin{bmatrix}\alpha_i^t\\ \beta_i^t\end{bmatrix} \tag{6}$$

其中，θ_i 为第 i 个量子比特位的旋转角。QEA 算法的步骤如下。

步骤 1　令 $t=0$，初始化量子种群 $Q(t)=\{q_1^t,q_2^t,\cdots,q_m^t\}$。

步骤 2　观测种群中的每个个体，使之由叠加态变为稳定态。针对个体 q_k 的第 i 个量子比特位，其概率幅表示为：

$$\begin{bmatrix}\alpha_k^i\\ \beta_k^i\end{bmatrix} \tag{7}$$

生成一个 0 到 1 之间的随机数 γ。如果 $\gamma<\alpha_k^{i2}$，则对应量子位取 0，否则取 1。因此，当 α_k^i 越大时，该量子比特位在观测时取 0 的概率就越大，反之该量子比特位取 1 的概率就越大。

步骤 3　根据步骤 2 的观测结果，对个体进行适应度评估，选取适应度最佳的个体作为当前的局部最优个体 $local_{best}$。将 $local_{best}$ 与全局最优个体 $global_{best}$ 进行比较，对全局最优个体进行更新。当满足算法终止条件时，终止算法，否则进行下一步。

步骤 4　根据观测结果，利用量子门对种群 $Q(t)$ 进行更新。

步骤 5　令 $t=t+1$，返回步骤 2。由于量子进化算法中的个体可以表示多种状态的叠加，使得较小规模的种群能够具备丰富的种群多样性。同时，量子门的引入也使得个体能够加速算法的收敛，找到最优解。

3　MSQEA

传统的采样方法在解决非平衡数据分类时主要有以下两个问题：1)现有方法在进行采样时具有一定的盲目性。在非平衡数据集中，对不同区域中的样本进行采样的效果也存在差异。传统的过采样方法在对边界区域的样本进行采样时，容易合成错误的少数类样本，而在对少数类样本的中心区域进行采样时也会合成冗余的样本，导致过拟合问题。欠采样由于盲目地移除多数类样本，因此容易丢失有价值的样本。

2)为了使多数类样本数量和少数类样本数量保持相对平衡，采用单一的采样方法对极度不平衡的数据集进行采样时需要合成或移除大量的样本，进而会出现合成过多冗余的样本或移除过多有效的样本的问题。

为了解决上述问题，首先采用量子进化算法找到合适的候选采样子集，然后对候选采样子集进行混合采样；其次，为了有效评价量子个体的适应度，使用聚类算法对原始数据集进行聚类，构建合适的验证集。

3.1　MSQEA 算法的介绍

在非平衡数据中，由于整个数据集样本的数量非常庞大，暴力搜索最优的候选采样子集并不是一个可行的方案。在量子进化算法中，每个个体在观测前可以表示为多种状态的叠加，而在观测后得到的 01 序列只能表示其中的一种状态，因此非常适合求解组合优化问题。采用量子进化算法对数量庞大的样本进行优化选择是一个较为合理的方案。

为了得到合适的候选采样集合，我们对多数类样本和少数类样本进行编码，构成量子的个体。每个个体可以分为两个部分 I_{maj} 和 $I_{\min}$，分别代表对多数类和少数类样本的编码结果，每个样本对应量子个体中的一个量子比特位。量子个体的编码示意图如图 1 所示。在对个体进行观测后会得到一组 0 和 1 组成的序列。对 I_{maj} 部分进行欠采样，将量子位为 0 对应的样本进行移除。对 $I_{\min}$ 部分进行过采样，对量子位为 1 对应的样本进行过采样。为了解决非平衡数据中的类别重叠和小析取问题，我们采用过采样方法 LADBMOTE 对 $I_{\min}$ 部分的样本进行过采样。在对样本进行混合采样后，利用采样后的平衡数据集使用分类算法进行训练，得到每个个体的分类器，然后利用验证集进行分类测试，将得到的分类评价指标值作为该个体的适应度。

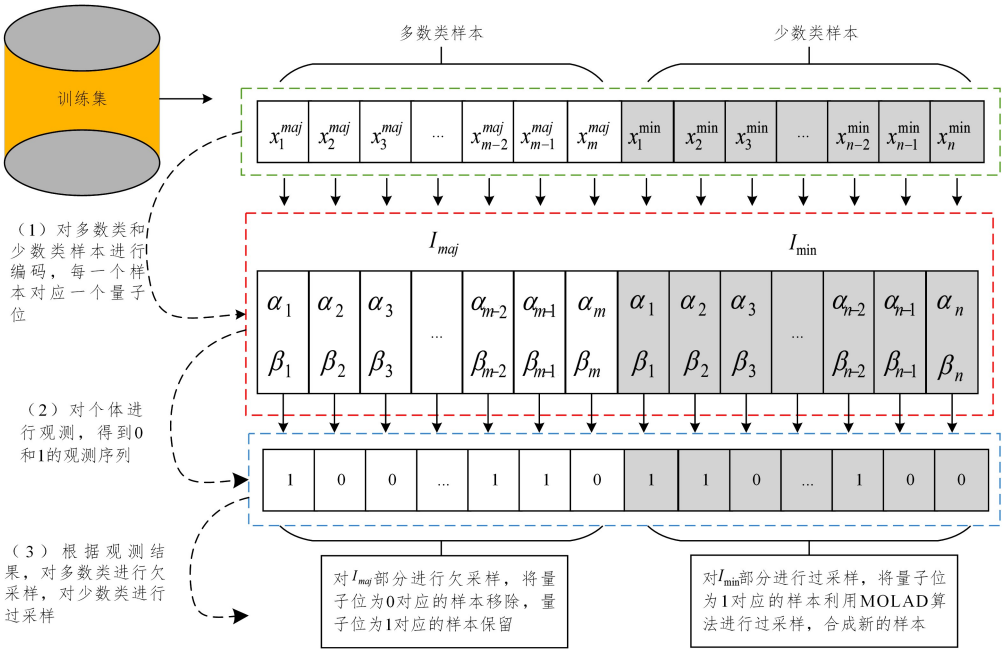


图 1　量子个体编码示意图

Fig. 1　Encoding chart of quantum individual

针对上文提到的问题(1)，MSQEA 使用量子进化算法选择出合适的候选采样子集，避免可能合成错误样本或移除有效样本的问题。针对问题(2)，根据个体的观测结果首先对多数类样本进行欠采样，从而在一定程度上缓解了数据集中类别不平衡的问题，然后利用 LADBMOTE 算法对选择后的少数类样本进行过采样，避免了合成大量冗余样本的问题。

3.2　基于聚类技术的验证集构建方法

在量子进化算法中，为了能够有效地评价每个个体在当前观测状态下的适应度，需要使用验证集对每个个体的分类器进行验证，将得到的分类评价指标值作为该个体的适应度。然后，根据适应度调整每个量子个体的旋转角度，从而使当前个体朝着最优个体进化。因此，验证集的好坏直接决定了整个量子进化算法中个体的寻优能力。如果直接使用测试集对分类器进行验证，必然会使最终的模型仅仅适用于测试集中的样本，导致泛化性能不足。而随机从训练集中选取样本作为验证集，每个个体的寻优能力受随机性的影响较大，一旦验证集的样本分布过于集中或分散，也会使得分类器无法学习到数据集的整体特征，从而出现欠拟合。因此，直接使用测试集或随机从训练集中选取样本作为验证集并不合适。验证集的样本分布必须能够反映整个原始数据集的样本分布，才能有效地评价个体的适应度。

基于聚类算法的思想，每个簇中的样本相似度高，而簇间的样本相似度低，因此每个簇中的聚类中心样本代表着该簇中的样本分布[5]。设定验证集的样本数量个数为 ε，然后将初始聚类中心数设置为验证集中样本的数量 ε。利用聚类算法对原始数据集进行聚类，将最终聚类得到的聚类中心样本作为验证集，从而使得验证集能够较好地代表整个原始数据集的样本分布，进而能够有效、准确地评价个体的适应度。MSQEA 算法的整体流程如图 2 所示。

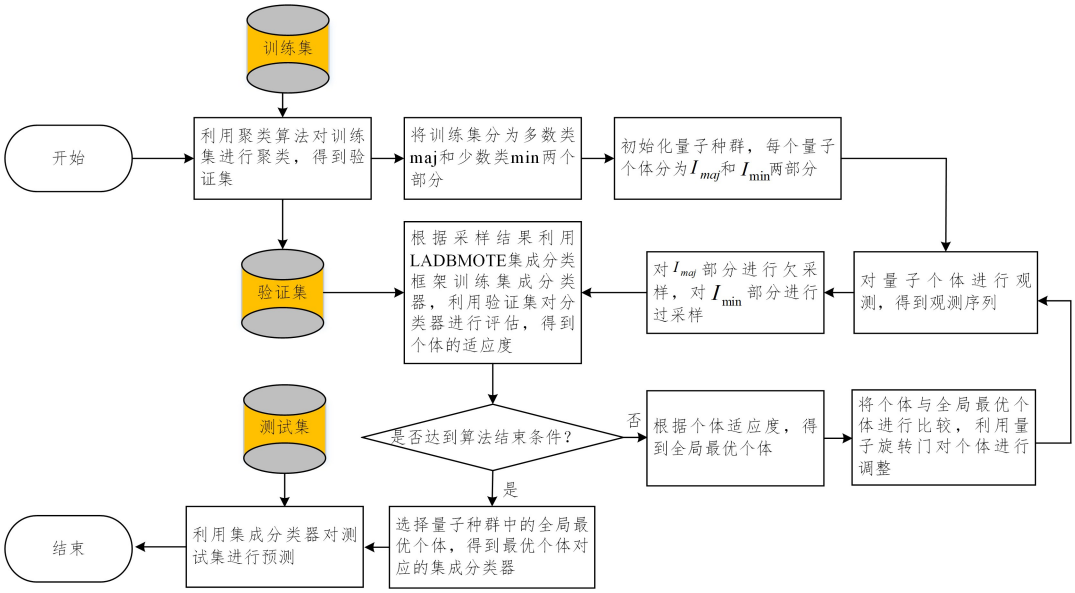


图 2 MSQEA 算法的流程图

Fig. 2 Flow chart of MSQEA algorithm

4 实验结果及分析

实验环境基于 JDK1.8 版本，在 WEKA 数据挖掘平台上做分类测试。为了验证 MSQEA 算法的有效性，本文从 KEEL 网站上下载了 6 个公开的非平衡数据集进行实验测试[20]。数据集的具体信息如表 1 所列。表 1 中，IR 为非平衡率，代表多数类与少数类样本数量的比值。为了保证算法的稳定性，文本采用五折交叉验证法，将每个数据集平均分为 5 份，保持每份数据集中非平衡率大致相同，依次使用其中一份作为测试集，其余 4 份作为训练集，分别运行 5 次，最终返回这 5 次分类结果的均值[21]。在数据集中，对比算法采用了 SMOTE，CBS 和 LADBMOTE 这 3 种较为优秀的采样算法。实验采用的分类算法有 SMO，J48 和 NB(Naïve Bayes)，分类器的参数均设置为 Weka 平台中的默认参数。

表 1 数据集

Table 1 Data set

Dataset	size	IR
pima	768	1.84
glass016v2	192	10.29
glass1	214	1.82
ecoli3	336	8.6
vehicle1	846	2.9
yeast1v7	459	14.3

在 MSQEA 中，每个个体利用验证集进行分类测试，将得到的分类评价指标值作为该个体的适应度。因此，选择合适的分类评价指标是评估个体质量的关键。传统的分类指标有分类精度 Accuracy、精确率 Precision 和召回率 Recall 等，这些指标只适用于平衡数据集的分类任务。在非平衡数据的分类任务中，常用的评价指标有 AUC，G-Mean 和 F-Measure 等。本文中，实验采用的评价指标为 AUC，因此在 MSQEA 中，为了使得算法能够取得最优的 AUC 值，本文使用 AUC 值作为每个个体的适应度。

4.1 实验参数对算法分类效果的影响

在 MSQEA 算法中，迭代次数、聚类比例和验证集的构建方式可能会对分类性能有所影响，以下分别分析它们的影响，以指导实验参数的设置。

(1)迭代次数对分类性能的影响

为了研究迭代次数对 MSQEA 采样效果的影响，结合实际可行性，将 MSQEA 中量子进化算法的迭代次数设定为[50，1 000]。在 MSQEA 处理后的数据集 pima，glass1 和 ecoli3 上做分类实验，分类器选用 J48，分类评价指标选用 AUC 值。实验结果如图 3 所示。

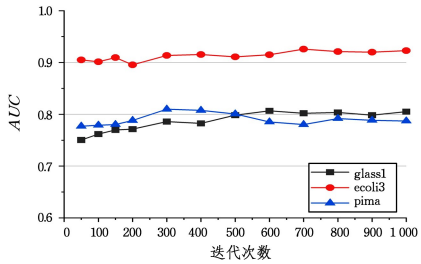


图 3 迭代次数对分类性能的影响

Fig. 3 Variation of classification performance for different iteration number

由图 3 可知，在 ecoli3 数据集上，分类性能受迭代次数的影响较小，AUC 值稳定在 0.9～0.92 之间。而在 pima 和 glass 数据集上，随着迭代次数的增加，分类性能逐渐提高，当迭代次数大于 400 时，分类性能基本维持稳定。综合分类性能和时间开销等因素，本文在后续实验中的迭代次数设置为 500。

(2)聚类比例对分类效果的影响

为了得到有效的验证集，MSQEA 使用聚类算法对原始数据集进行聚类，然后将聚类中心样本作为验证集，因此聚类中心的个数直接影响着验证集的质量。若设置的聚类中心数太小，将导致验证集无法有效地反映整个数据集的样本分布；

若聚类中心数设置得太大，则可能会降低聚类效果。为了研究聚类中心数对算法效果的影响，本文设定聚类中心数占数据集样本总数的比例范围为[0.2，0.9]，步长设置为 0.1，在数据集 glass1 上的分类性能如图 4 所示。

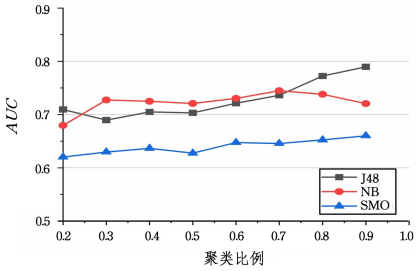


图 4　聚类比例对分类性能的影响

Fig. 4　Variation of classification performance for different clustering ratio

由图 4 可知，当使用 J48 和 SMO 分类器时，随着聚类比例的增大，分类性能逐渐提升。而使用 NB 分类器时，聚类比例在 0.3～0.7 之间可以取得较好的效果。当聚类比例大于 0.7 时分类性能有所下降，但是 AUC 值仍能维持在 0.7 以上。因此，综合 MSQEA 算法在这 3 种分类器上的效果，本文在后续实验中将聚类比例参数设置为 0.8。

(3) 验证集的构建方式对算法分类效果的影响

MSQEA 利用聚类算法对原始数据集进行聚类，构建一个适合的验证集，使得验证集能够较好地反映原始数据集的分布。相比随机从数据集中选择样本构成验证集，使用聚类算法构建验证集能够更好地避免过拟合或欠拟合问题。为了验证 MSQEA 中使用聚类技术构建验证集的有效性，MSQEA 分别采用随机从数据集中选择样本构建验证集和使用聚类技术构建验证集这两种方法进行对比实验。RS(Random Selection)方法表示随机从数据集中选择样本作为验证集，CS(Clustering Selection)方法表示使用聚类技术构建验证集。由于 RS 方法具有较大的随机性，因此，为了消除随机性对实验结果的影响，RS 方法分别运行 5 次，取平均值作为最终的分类结果。两种方法的分类效果如表 2 所列。

表 2　以 RS 和 CS 构建验证集的分类效果

Table 2　Classification performance using RS and CS to build validation set

	J48		SMO		NB	
	RS	CS	RS	CS	RS	CS
pima	0.785	**0.787**	**0.756**	0.752	0.826	0.826
glass016v2	0.728	**0.757**	0.655	**0.761**	0.689	**0.747**
glass1	0.797	0.795	0.632	**0.623**	0.722	**0.746**
ecoli3	0.919	0.916	0.894	**0.901**	0.926	**0.946**
vehicle1	0.815	0.814	0.771	**0.786**	0.749	**0.753**
yeast1v7	0.776	**0.788**	0.776	**0.809**	0.783	**0.807**
Avg	0.803	**0.810**	0.747	**0.772**	0.783	**0.804**

由表 2 可知，使用 CS 方法构建验证集的平均分类效果在 6 个数据集上均优于使用 RS 方法构建验证集的平均分类效果。当使用 J48 分类器时，使用 CS 方法的分类效果与使用 RS 方法时差距较小，整体提高了约 0.9%。当使用 SMO 和 NB 分类器时，使用 CS 方法的分类效果相比 RS 方法分别提高了 3.3%和 2.7%。因此，CS 方法更适用于 SMO 和 NB 分类器。

4.2　算法性能分析

为了验证本文提出的 MSQEA 算法的有效性，本文与经典过采样算法 SMOTE、欠采样算法 CBS 和过采样算法 LADBMOTE 分别在 3 种不同的分类器上进行比较，分类评价指标为 AUC 值。量子进化算法的迭代次数为 500，聚类比例为 0.8。实验结果如表 3－表 5 所列。

表 3　以 J48 为分类器的实验结果

Table 3　Classification performance on J48

Data set	SMOTE	CBS	LADBMOTE	MSQEA
pima	0.735	0.777	**0.807**	0.787
glass016v2	0.608	0.7	0.709	**0.757**
glass1	0.724	0.709	**0.797**	0.795
ecoli3	0.83	0.880	0.902	**0.916**
vehicle1	0.708	0.766	**0.817**	0.814
yeast1v7	0.671	0.786	0.777	**0.788**

表 4　以 SMO 为分类器的实验结果

Table 4　Classification performance on SMO

Data set	SMOTE	CBS	LADBMOTE	MSQEA
pima	0.746	0.740	**0.756**	0.752
glass016v2	0.596	0.609	0.650	**0.761**
glass1	0.564	0.595	0.575	**0.623**
ecoli3	0.886	0.867	0.897	**0.901**
vehicle1	0.754	0.734	0.770	**0.786**
yeast1v7	0.757	0.755	0.778	**0.809**

表 5　以 NB 为分类器的实验结果

Table 5　Classification performance on NB

Data set	SMOTE	CBS	LADBMOTE	MSQEA
pima	0.723	0.814	0.817	**0.826**
glass016v2	0.743	0.646	0.656	**0.747**
glass1	0.543	0.688	0.676	**0.746**
ecoli3	0.906	0.940	0.934	**0.946**
vehicle1	0.621	0.719	0.716	**0.753**
yeast1v7	**0.828**	0.802	0.802	0.807

由表 3 可知，当使用 J48 分类器时，MSQEA 算法在数据集 glass016v2，ecoli3 和 yeast1v7 上相比其余 3 种对比算法均能得到最优的分类性能。在 glass016v2 数据集上，MSQEA 的 AUC 值比 SMOTE，CBS 和 LADBMOTE 分别提高了 24.5%，8.1%和 6.8%；在 ecoli3 数据集上，MSQEA 的 AUC 值比 SMOTE，CBS 和 LADBMOTE 分别提高了 10.3%，4.0%和 1.6%；在 yeast1v7 数据集上，MSQEA 的 AUC 值比 SMOTE，CBS 和 LADBMOTE 分别提高了 17.4%，0.3%和 1.4%。

由表 4 可知，当使用 SMO 作为分类器时，MSQEA 的分类性能仅在数据集 pima 上次于 LADBMOTE 算法，在其他数据集上均优于其余 3 种对比算法。在 glass016v2 数据集上，MSQEA 的 AUC 值比 SMOTE，CBS 和 LADBMOTE 分别提高了 28%，25%和 17%；在 glass1 数据集上，MSQEA 的 AUC 值比 SMOTE，CBS 和 LADBMOTE 分别提高了 10%，4.7%和 8.3%；在 ecoli3，vehicle1 和 yeast1v7 数据集上，MSQEA 算法的 AUC 值相比其余 3 种算法平均分别提升了约 4.2%，6.2%和 2.1%。

由表 5 可知，当使用 NB 作为分类器时，MSQEA 的分类性

能仅在数据集 yeast1v7 上次于 SMOTE；在 pima，glass016v2，glass1，ecoli3 和 vehicle1 上，MSQEA 算法的 AUC 值相比 SMOTE，CBS 和 LADBMOTE 平均分别提升了约 15.5%，6.2%和 6.3%。

上述实验结果表明，MSQEA 的采样效果优于 SMOTE，CBS 和 LADBMOTE，在经过 MSQEA 处理的数据集上，J48，SMO 和 Naïve Bayes 分类器的分类性能均有较大的提升。

结束语 为了解决单一采样方法在解决数据非平衡分布时容易造成过拟合或丢失重要样本的问题，本文提出了一种基于量子进化算法的混合采样方法 MSQEA。该方法利用量子进化算法选择合适的采样子集，将基于样本局部密度的过采样方法 LADBMOTE 和欠采样结合，对采样子集进行混合采样。为了有效地评价量子个体的适应度，使用聚类算法构建合适的验证集对个体进行验证，得到个体的适应度。在多种分类算法下的实验结果表明，本文提出的 MSQEA 算法的采样效果优于单一的采样算法，经过 MSQEA 算法处理的数据集在多种分类器上的分类性能均有较大提高。然而，MSQEA 仅解决了二分类非平衡数据问题，如何针对多类别非平衡数据进行采样将是今后的研究重点。

参 考 文 献

[1] SUN A, LIM E P, LIU Y. On strategies for imbalanced text classification using SVM: A comparative study[J]. Decision Support Systems, 2009, 48(1): 191-201.

[2] MAZUROWSKI M A, HABAS P A, ZURADA J M, et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance [J]. Neural networks, 2008, 21(2-3): 427-436.

[3] CAO H, LI X L, WOON D Y K, et al. Integrated oversampling for imbalanced time series classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(12): 2809-2822.

[4] DHEEPA V, DHANAPAL R, MANJUNATH G. Fraud detection in imbalanced datasets using cost based learning[J]. Eur. J. Sci. Res, 2012, 91: 486-490.

[5] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data [J]. Information Sciences, 2017, 409: 17-26.

[6] BARUA S, ISLAM M M, YAO X, et al. MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405-425.

[7] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[8] ZHU T, LIN Y, LIU Y. Synthetic minority oversampling technique for multiclass imbalance problems[J]. Pattern Recognition, 2017, 72: 327-340.

[9] YANG H, CHEN H M. Ensemble classification algorithm for imbalanced data combining the local area density[J]. Journal of Frontiers of Computer Science and Technology. 2020, 14(2): 274-284.

[10] CANO J R, HERRERA F, LOZANO M. Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study[J]. IEEE Transactions on Evolutionary Computation, 2003, 7(6): 561-575.

[11] AHA D W, KIBLER D, ALBERT M K. Instance-based learning algorithms[J]. Machine Learning, 1991, 6(1): 37-66.

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

[13] TSAI C F, LIN W C, HU Y H, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019, 477: 47-54.

[14] SHAO K, ZHAI Y, SUI H, et al. Learning from the imbalanced data based on quantum evolutionary[J]. ICIC Express Letters, 2014, 8(6): 1725-1729.

[15] LI J, FONG S, WONG R K, et al. Adaptive multi-objective swarm fusion for imbalanced data classification[J]. Information Fusion, 2018, 39: 1-24.

[16] WU Y F, LIANG J Y, WANG J H. Classification algorithm based on hybrid sampling for unbalanced data[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(2): 342-349.

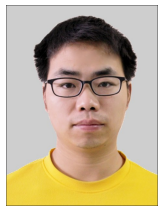
[17] HU F, WANG L, ZHOU Y, et al. An oversampling method for imbalance data based on three-way decision model[J]. Acta Electronica Sinica, 2018, 46(1): 135-144.

[18] 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. Springer, Berlin, Heidelberg, 2005: 878-887.

[19] HAN K H, KIM J H. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization[J]. IEEE Trans on Evolutionary Computation, 2002, 6(6): 580-593.

[20] ALCALÁ-FDEZ J, FERNÁNDEZ A, LUENGO J, et al. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework[J]. Journal of Multiple-Valued Logic & Soft Computing, 2011, 17: 255-287.

[21] MORENO-TORRES J G, SÁEZ J A, HERRERA F. Study on the impact of partition-induced dataset shift on k-fold cross-validation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(8): 1304-1312.



YANG Hao, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include database technology and data mining.

CHEN Hong-mei, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include granular calculation, rough sets and intelligent information processing.