第 35 卷　第 22 期　　计　算　机　工　程　　2009 年 11 月
Vol.35　No.22　　Computer Engineering　　November 2009
·人工智能及识别技术·　文章编号：1000—3428(2009)22—0182—03　文献标识码：A　中图分类号：TP391

基于聚类和遗传交叉的少数类样本生成方法

杜　娟，衣治安，周　颖
(大庆石油学院计算机与信息技术学院，大庆 163318)

摘　要：传统的分类算法在处理不均衡样本数据时，分类器预测倾向于多数类，样本数量少的类别分类误差大。针对该问题，提出一种基于聚类和遗传交叉的少数类样本上采样方法，通过 K-means 算法将少数类样本聚类分组，在每个聚类内使用遗传交叉获取新样本，并进行有效性验证。基于 K-最近邻及支持向量机分类器的实验结果证明了该方法的有效性。

关键词：不均衡数据集；分类；聚类；遗传交叉

Generation Method for Samples of Minority Class Based on Clustering and Genetic Crossover

DU Juan, YI Zhi-an, ZHOU Ying
(Institute of Computer and Information Technology, Daqing Petroleum Institute, Daqing 163318)

【Abstract】Prediction results of classification with traditional classify algorithm are towards the class with more samples when training imbalanced data sets. The classification error of the minority class is grave. Aiming at the problem, this paper proposes an over-sampling method based on clustering and genetic crossover. The samples of minority class are grouped by using K-means clustering algorithm. Genetic crossover algorithm is used in each cluster to gain new samples and confirm the validity. The validity of the method is proved through the experiments of K-Nearest Neighbor(KNN) and Support Vector Machine(SVM) classification.

【Key words】 imbalanced data set; classification; clustering; genetic crossover

不均衡数据集的分类问题是机器学习和模式识别领域的研究热点，许多实际应用中都存在不均衡数据集，如医疗诊断、入侵检测、信息过滤、文本分类。不均衡数据集是指在同一样本集合中某些类别较其他类别在样本数量上占据较大的优势。容易获取样本的类称为多数类，反之称为少数类。对于不均衡数据集上的分类，往往少数类的错分代价更大。例如使用文本分类的方法实现有害信息过滤时，包含有害信息的样本不易获取，造成正、负面样本分布严重不均衡，然而人们更加关注的是少数类的过滤性能，大量有害文本被错分为正常文本是不能容忍的。

1　不均衡数据集对分类精度的影响

传统的分类算法如 K-最近邻(K-Nearest Neighbor, KNN)、支持向量机(SVM Support Vector Machine, SVM)都是以类分布基本平衡为前提假设，在不均衡数据集上的分类效果严重倾向于多数类样本。KNN 是一种典型的基于类比的学习算法，因此，每个类别都必须具有一定数量及代表性的训练样本才能保证分类的准确性。文献[1]分析了不均衡样本数据对 SVM 分类精度的影响，同时证明了当训练样本数量趋于平衡时，SVM 预测倾向性会急剧减小。目前对于该问题的解决方法可以分为移动阈值法、调整代价或权重法和采样法。本文主要讨论第 3 种方法。

采样法属于数据层的方法，通过对多数类的下采样或少数类的上采样来降低不平衡性，从而提高少数类的分类性能。最简单的下采样就是随机地去掉某些样本，减小多数类样本的规模使各类的分布趋于平衡，但是会导致丢失多数类的重要信息。最简单的上采样方法是单纯地复制少数类样本，增加其数量，但是由于没有引入新的类别信息，因此分类器学到的决策域变小，导致过学习。对此，研究者们提出了一些改进的方法，如文献[2]提出的通过采样技术合成少数点移出冗余多数点的智能方法、文献[3]提出的基于初分类的过抽样算法，都能在一定程度上提高少数类的分类性能。

本文提出了一种基于聚类和遗传交叉的少数类样本生成方法，属于上采样方法。实验结果表明此方法能有效提高少数类样本分类性能。

2　聚类和遗传算法

2.1　聚类

聚类是按照数据或对象的某些属性将其聚集成若干个聚类(簇)，使同一个簇内的数据或对象尽可能相似，不同簇间的差异尽可能大。聚类分析可以作为一种获得数据分布情况、观察类特征并做进一步分析的工具。典型的基于划分的聚类方法是 K-means 方法，算法描述如下：

(1)假设需要聚类的类为 $\boldsymbol{C}_P$，输入簇的个数 k 和样本集。

(2)从类 $\boldsymbol{C}_P$ 的样本集中随机选取 k 个向量作为初始的聚类中心：$\boldsymbol{C}_P=\{\boldsymbol{C}_1^P,\boldsymbol{C}_2^P,\cdots,\boldsymbol{C}_k^P\}$。

(3)将待分类的样本 x_i，$i=1,2,\cdots,n$ 分配给某一个聚类中心 $\boldsymbol{C}_j^p$，$1<j<m$，决策依据是 $\|x_i-\boldsymbol{C}_j\|=\min\limits_{1\le j\le m}\|x_i-\boldsymbol{C}_j\|$。

基金项目：黑龙江省研究生创新科研资金资助项目(YJSCX2006-38HLJ)

作者简介：杜　娟(1980－)，女，讲师、硕士，主研方向：人工智能，数据挖掘；衣治安，教授；周　颖，硕士

收稿日期：2009-03-02　**E-mail：**dqpidj@163.com

(4)计算 m 个簇中样本的均值向量：$\boldsymbol{C}_j = 1/N_j \sum_{x \in s_j} x$，$j = 1,2,\cdots,m$，其中，$N_j$ 是第 j 个聚类域 S_j 中的样本数；$\boldsymbol{C}_j$ 作为新的第 j 个簇的聚类质心，形成新的聚类中心 $\boldsymbol{C}_p^{'}$。

(5)若聚类中心不再变化，终止算法，输出新的簇集合；否则，转步骤(3)。以上迭代过程采用下面的目标函数进行：$J = \min(\sum_{i=1}^{n} \sum_{x \in s_j} |x_i - \boldsymbol{C}_j|^2), j = 1,2,\cdots,k$，其中，$s_j$ 是中心为 $\boldsymbol{C}_j$ 的聚类域。

2.2 聚类效果的评价

聚类效果的评价原则是：聚类内部的对象尽可能相似，聚类之间尽可能远离。本文使用聚类中心距离矩阵和聚类域中样本与聚类中心的距离方差 2 个指标来衡量聚类效果[4]。初始中心点以及 k 值的选择也会影响聚类结果，通过实验对比，当 k=5 时，样本距离方差达到最小值。

2.3 遗传算法和交叉算子

2.3.1 遗传算法

遗传算法(Genetic Algorithms, GA)是一种借鉴生物界自然选择和遗传机理的进化算法。遗传算法编码方式分为二进制编码(编码符号集合为{0,1})和实数编码。交叉操作是遗传算法区别于其他优化算法的根本所在，是产生下一代的重要操作。交叉又称为重组，是以较大的概率从群体中选择 2 个个体，交换个体的某个或某些位，交叉操作的关键是交叉算子的设计，与实际的问题和编码方式有关。适合二进制编码方式的交叉算子有单点交叉、两点交叉、多点交叉、均匀交叉等。

2.3.2 GA 实数编码方式的交叉算子

实数编码中染色体的基因是由某个范围内的浮点数表示的。其交叉算子主要包括离散交叉、算术交叉、线性交叉和启发式交叉等。本文采用的是线性交叉算子[5]，新个体是 2 个父类染色体的线性组合，一次交叉后可产生 3 个子代个体。线性交叉产生的子代个体不过分亲近于父代，同时不会产生大量的无效基因，具体表示如下：

$$\begin{cases} \boldsymbol{Y}_1 = 1.5\boldsymbol{X}_1 - 0.5\boldsymbol{X}_2 \\ \boldsymbol{Y}_2 = -0.5\boldsymbol{X}_1 + 1.5\boldsymbol{X}_2 \\ \boldsymbol{Y}_3 = (\boldsymbol{X}_1 + \boldsymbol{X}_2)/2 \end{cases} \quad (1)$$

其中，$\boldsymbol{X}_1, \boldsymbol{X}_2$ 为父类个体向量；$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3$ 为新个体向量。假设个体的变量个数为 2，则线性交叉后个体的可能位置如图 1 所示。

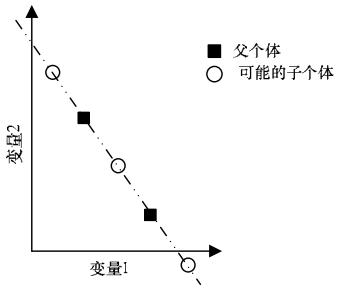

图 1 线性交叉个体位置

3 基于聚类和遗传算法的样本生成方法

受聚类和遗传算法的启发，可以使用聚类描述少数类样本空间内部结构，将相似样本聚集成簇，在簇内采用 GA 线性交叉算子生成新样本。这样，新样本在特征尽可能相似的样本间交叉产生，同时各个簇中均获得一定比例的新样本，能够保证新样本具有更好的覆盖性和代表性，更加符合原样本空间的分布。

3.1 文本的表示

数学表示是文本分类的基础[6]，目前应用比较普遍的文本表示模型是向量空间模型 VSM。在 VSM 中文档被表示成向量 $\boldsymbol{X}_j = (w_1^j, w_2^j, \cdots, w_n^j), 1 \le j \le N$，其中，$N$ 为文档集中包含的文档数量；w_1^j 为文档 j 第 i 个特征项的权重。关于权重的计算有很多方法，本文采用了一种比较普遍的 TFIDF 公式：

$$w_t^d = \frac{tf_t^d \times \log_n(N/n_t + 0.01)}{\sqrt{\sum_{t \in d}[tf_t^d \times \log_n(N/n_t + 0.01)]^2}} \quad (2)$$

其中，w_t^d 为词 t 在文本 d 中的权重，是取值为 0~1 的实数，对应遗传算法的实数编码方式；tf_t^d 为 t 在文本 d 中的词频；n_t 为样本集中包含 t 的文本数；分母为归一化因子。

3.2 新样本的生成

先使用 K-means 方法将样本空间划分为特征相似的聚类。再使用线性交叉算子生成新的个体，新个体将继承父代个体的特征，但并不是单纯的复制。最后按照一定的策略对新个体进行筛选，抛弃无效样本。算法具体步骤如下：

(1)输入 k=5，使用 K-means 算法聚类将类 $\boldsymbol{C}_P$ 划分为 5 个聚类，第 i 个聚类表示为 $\boldsymbol{C}_i^p, 1 \le i \le 5$。

(2)在 $\boldsymbol{C}_i^p$ 内部对 m_i 个体进行随机配对，配对组数为 $m_i/2$。

(3)对每对个体按 0.8 的概率进行线性交叉操作，保留全部子代及父代个体。则经过一轮交叉后，$\boldsymbol{C}_p$ 中样本数最多将达到 $2.5m_i$ 个，规模是原样本空间的 2.5 倍。

(4)验证 $\boldsymbol{C}_i^p$ 中新个体的有效性，抛弃无效样本。

(5)如果样本数量没有达到要求，转步骤(2)进行下一轮交叉，直至获得要求数量的样本；否则，算法结束，输出 $\boldsymbol{C}_p$ 作为新的样本空间。

步骤(4)中的有效性验证是指，通过子代样本的聚类质心距离来控制样本的质量。对于新样本，总是期望其具有随机性并且保留较多的类别信息，即尽可能具有好的代表性，但仍然会有一些新样本是不适合的，超出了聚类的范围，称其为无效样本，直接抛弃。选择策略涉及的定义如下：

假设类 $\boldsymbol{C}_p$ 的聚类质心为 $\boldsymbol{C}_p = \{\boldsymbol{C}_1^p, \boldsymbol{C}_2^p, \cdots, \boldsymbol{C}_m^p\}$。

定义 1 聚类 $\boldsymbol{C}_i^p$ 的半径为：$r_i^p = \max(\|x_j^p - \boldsymbol{C}_i^p\|), 1 \le j \le N_p$，$1 \le i \le m$，其中，$N_p$ 为 $\boldsymbol{C}_i^p$ 中样本的个数。

定义 2 样本 x_i 的聚类质心距离为：$d_{ji}^p = \|x_j^p - \boldsymbol{C}_i^p\|$，$1 \le i \le N_p, 1 \le j \le m$。

有效性验证的具体步骤如下：

(1)计算聚类 $\boldsymbol{C}_i^p$ 中新样本 Y_j 的聚类质心距离 d_{ji}^p。

(2)以聚类半径 r_i^p 为阈值，满足条件 $d_{ji}^p \le r_i^p$ 的新样本 Y_j 为有效样本，存入训练集合中。

4 实验及结果分析

本文从标准的 UCI 数据集中选取了 Breast Cancer 和 Vehicle 这 2 个数据集用于训练和测试 SVM 分类器；另外选择了 Sonar 和 Pima-Indian diabetes 集合用于训练和测试 KNN 分类器。其中，Breast Cancer, Vehicle 和 Pima-Indian diabetes 数据集中 2 类样本不均衡；Sonar 数据集中 2 类样本基本均衡，本文在 Sonar 数据集的 97 个正例样本中抽取 30 个样本

与负例样本构成不均衡数据集 Sonar-1。在这些数据集中随机选取 70%的样本作为训练集，其余作为测试样本集。表 1 为数据集的基本信息。

表 1　数据集基本信息

数据集名称	Breast Cancer	Vehicle	Sonar-1	Pima -Indian diabetes
特征数	30	18	60	8
多数类样本(所占比例)	458 (65.5%)	647 (76.5%)	111 (78.7%)	500 (65.1%)
少数类样本(所占比例)	241 (34.5%)	199 (23.5%)	30 (21.3%)	268 (34.9%)
训练样本数	489	592	99	538
测试样本数	210	254	42	230

实验过程中，SVM 的核函数选择径向基核函数，参数 C =2 000(惩罚因子)，δ =0.01(核参数)，遗传交叉的概率为 0.8，KNN 分类器的 K 值取 11。对于分类效果的评价，本文采用了 *F-value* 作为评价标准，它是综合了查全率 *Recall* 和查准率 *Precision* 的评价指标，只有当两者都较高时，才能获得较高的 *F-value*。*F-value* 的计算公式如下：

$$F-value=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Precision+Recall} \tag{3}$$

其中，参数 β 的值是可调的，通常取值为 1，最终统计 10 次实验少数类样本 *F-value* 平均值。表 2 和表 3 分别为 SVM 及 KNN 分类器在原始数据集及新数据集上的实验结果对比。分析可得，使用不同分类器分类后的均衡数据集的少数类分类效果都较好，KNN 分类器在 Sonar-1 数据集上 *F-value* 值提高较大，主要是由于 KNN 是典型基于类比的方法，少数类训练样本数量的增加对其产生了较大的影响。

表 2　SVM 分类器实验结果

	数据集	Breast Cancer	Vehicle
原始数据集	少数类样本	169	139
	多数类样本	320	453
	平均 *F-value*/(%)	92.19	90.05
本文方法调整后的样本集合	新生成样本	168	293
	新少数类样本(扩充比例)	337(200%)	432(300%)
	多数类样本	320	453
	平均 *F-value*/(%)	94.86	93.24
	F-value 提高率/(%)	2.67	3.19

表 3　KNN 分类器实验结果

	数据集	Sonar-1	Pima –Indian diabetes
原始数据集	少数类样本	21	188
	多数类样本	78	350
	平均 *F-value*/(%)	66.15	65.84
本文方法调整后的样本集合	新生成样本	55	156
	新少数类样本(扩充比例)	76(360%)	344(180%)
	多数类样本	78	350
	平均 *F-value*/(%)	74.85	71.26
	F-value 提高率/(%)	8.70	5.42

5　结束语

本文从对数据采样方法的角度出发，提出了基于聚类和遗传交叉的少数类样本生成方法。对 UCI 标准数据集以及本文方法处理后的新数据集进行了实验，结果表明，在新的分布均衡的训练集上，KNN 及 SVM 分类器都能获得更好的少数类分类效果。本文方法对于一些关注少数类分类精度的实际应用有重要的意义。但本文方法仅考虑了样本数量不均衡造成的分类问题，实际上样本分布的均匀程度对分类效果也有一定的影响，这是下一步将要解决的问题。

参考文献

[1] 吴洪兴, 彭　宇, 彭喜元. 适用于不平衡样本数据处理的支持向量机方法[J]. 电子学报, 2006, 34(12): 2395-2398.

[2] Chawla N, Bowyer K, Hall L, et al. Smote: Synthetic Minority over Sampling Technique[J]. Artificial Intelligence Research, 2002, 16: 321-356.

[3] 韩　慧, 王　路, 温　明, 等. 不均衡数据集学习中基于初分类的过抽样算法[J]. 计算机应用, 2006, 26(8): 1894-1897.

[4] 夏　锋, 彭　鑫, 赵文耘. 基于聚类方法的审计分层抽样算法研究[J]. 计算机应用与软件. 2008, 25(1): 14-16.

[5] Wright A H. Genetic Algorithms for Real Parameter Optimization. Foundations of Genetic Algorithms[M]. [S. l.]: Morgan Kaufmann, 1991.

[6] 黄　旭, 朱艳琴, 罗喜召. 实时文本分类系统的研究与实现[J]. 计算机工程, 2008, 34(18): 87-88, 92.

编辑　张　帆

(上接第 172 页)

5　结束语

本文提出直接在当前输入图像和全景图像间进行配准的算法，可以有效避免局部配准矩阵连乘造成的误差积累与传播问题。为克服摄像机二维多幅拍摄方式下图像拼接产生的不一致问题，提出了时间相邻特征点和空间相邻特征点共同配准的策略，有效克服了一般全局配准方法实时性差的问题。该视频图像直接拼接算法特别适合基于视频的大面积静态场景的实时观测。

参考文献

[1] Candocia F M. Jointly Registering Images in Domain and Range by Piecewise Linear Comparametric Analysis[J]. IEEE Transaction on Image Processing, 2003, 12(4): 409-419.

[2] Zhu Zhigang, Xu Guangyou, Edward M R. Fast Construction of Dynamic and Multi-resolution 360° Panoramas from Video Sequences[J]. Image and Vision Computing, 2006, 24(1): 13-26.

[3] Steve H, Harpreet S S, Rakesh K. Automated Mosaics via Topology Inference[J]. IEEE Computer Graphics and Applications, 2002, 22(2): 44-54.

[4] Kang E Y, Cohen I, Medioni G. A Graph-based Global Registration for 2D Mosaics[C]//Proceedings of the 15th International Conference on Pattern Recognition. Barcelona, Spain: [s. n.], 2000: 257-260.

[5] Gracias N, Jose S V. Underwater Video Mosaics as Visual Navigation Maps[J]. Computer Vision and Image Understanding, 2001, 79(1): 66-97.

[6] Hsu C T, Tsan Y C. Mosaics of Video Sequences with Moving Objects[J]. Signal Processing: Image Communication, 2004, 19(1): 81-98.

编辑　任吉慧