计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):223−226 [doi: 10.15888/j.cnki.csa.005911]

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

基于拆分集成的不均衡数据分类算法①

杜红乐, 张　燕

(商洛学院 数学与计算机应用学院, 商洛 726000)

摘　要: 为改进 SVM 对不均衡数据的分类性能, 提出一种基于拆分集成的不均衡数据分类算法, 该算法对多数类样本依据类别之间的比例通过聚类划分为多个子集, 各子集分别与少数类合并成多个训练子集, 通过对各训练子集进行学习获得多个分类器, 利用 WE 集成分类器方法对多个分类器进行集成, 获得最终分类器, 以此改进在不均衡数据下的分类性能. 在 UCI 数据集上的实验结果表明, 该算法的有效性, 特别是对少数类样本的分类性能.

关键词: 支持向量机; 不均衡数据集; 分类器集成

引用格式: 杜红乐,张燕.基于拆分集成的不均衡数据分类算法.计算机系统应用,2017,26(8):223–226. http://www.c-s-a.org.cn/1003-3254/5911.html

Imbalanced Data Classification Algorithm Based on Split and Classifier Ensemble

DU Hong-Le, ZHANG Yan

(School of Mathematics and Computer application, Shangluo University, Shangluo 726000, china)

Abstract: To improve the performance of Support Vector Machine classifier for imbalanced data, an imbalanced data classification algorithm based on split and classifier ensemble is introduced. The majority class sample is divided into several sub sets by clustering, and each subset is combined with minority class sample to produce a training subset. Then the training subsets are learned and multiple classifiers are obtained. Finally the multiple classifiers are integrated and the ensemble classifier is obtained. Experimental results show the algorithm is effective for imbalanced dataset, especially for the minority class samples.

Key words: support vector machine; imbalanced dataset; classifier ensemble

引言

传统支持向量机算法多是针对样本错分代价相同的均衡数据, 而现实中的不均衡数据广泛存在, 针对不均衡数据的分类问题是机器学习研究的一个热点. 不均衡数据广泛存在于各个领域, 而在不均衡数据的分类中, 更关注少数类的分类准确率, 如入侵检测、医疗诊断、故障检测等, 研究者对该分类问题做了大量研究, 主要从数据层面[1-3]和算法层面[4-7]给出相应的解决方案, 主要目标是提高少数类的分类准确率.

当前, 不均衡数据分类常采用的方法有数据重采样[1-3]、代价敏感[4]、单分类[5]、集成学习[6-8]等, 各种方法存在各自的不足, 重取样算法的过取样, 会出现过拟合现象, 欠取样可能会丢掉很多有用的空间信息, 导致对大类的欠学习; 代价敏感算法对不同类设定不同的错分代价, 但错分代价依据经验估计或者依据历史数据进行预测, 都无法准确描述不均衡数据集对决策函数的影响; 集成学习通过迭代计算样本的权重及分类器的权重, 使得对不均衡数据. 针对以上不足本文提出一种拆分集成的不均衡数据分类算法(Split and Classifier Ensemble for Imbalanced data, SCE-SVM), 该算法首先对多少类样本安装与少数类样本的比例进行聚类, 然后从每个簇中选择样本与少数类样本一起构成训练子集, 然后采用 AdaBoost 算法训练子分类器并集成, 该方法可以避免过取样的过拟合现象, 也可以避免欠取样中丢掉有用的样本信息. 仿真实验表明该算法对不均衡数据有较好的分类准确率, 特别是少数类样本的分类准确率.

① 基金项目: 陕西省自然科学基础研究计划(2015JM6347); 陕西省教育厅科技计划(15JK1218); 商洛学院科学与技术研究项目(15sky010)
收稿时间: 2016-12-12; 采用时间: 2017-01-04

1 相关工作

1.1 K 均值聚类算法

K-Means 算法是一种应用广泛的、经典的聚类算法, 依据聚类后的簇数 K 进行聚类, 对于事先知道类别数的数据有较好的聚类效果. 对不均衡数据的处理中, 常用的重取样有欠取样和过取样, 如何使得训练集中的类样本达到均衡, 并且避免欠取样和过取样的不足, 是针对不均衡数据重取样方法的出发点. 林舒杨等[9]利用 K-Means 算法对多数类样本进行聚类并提取簇中心, 用所有簇中心构成新的多数类样本, 与少数类样本重构新的训练集, 使得多数类与少数类之间达到均衡, 但是选取每个簇中心的同时, 丢掉了簇中其它样本包含的信息, 造成对原始训练集的欠学习.

1.2 集成分类器

集成分类器的性能被 Wang[10]等从理论上证明优于单个分类器的性能, 基于权重的集成分类器(Weight Ensemble Classifiers, WE)被普通认为是简单而有效的集成方法. 给定数据子集$T_i, i=1,2,\cdots,m$, WE 集成分类器采用一种算法(本文选择 SVM)对各子集进行学习, 得到 m 个子分类器 f_i, 然后采用某种权重计算方法对每个分类器赋予权重 w_i, 最后得到集成分类器为:

$$f(x)=\sum_{i=1}^{m} w_i f_i \tag{1}$$

其中$\sum_{i=1}^{m} w_i=1, 0\le w_i\le 1$.

对测试集中的数据, 认为与训练集中的数据有相同的分布, 采用集成分类器进行预测, 每个子分类器的权重计算后面会给出详细的计算方法.

2 基于拆分集成的不均衡数据分类算法

2.1 算法思想

针对过取样和欠取样存在的问题, 本文提出一种拆分集成的不均衡数据分类算法, 本算法在既不删除样本又不增加样本的基础上, 通过对多数类样本进行拆分, 所得子集与少数类合并构成各训练子集, 该方法可以充分利用少数类样本信息, 使得每个子集的“多数类”样本能保持原有的空间信息, 同时不会丢掉多数类样本中的任何信息. 由于多个分类器进行集成的分类器性能要优于单个分类器的性能, 本文采用 WE 集成方法对多个子集的分类器进行集成, 每个子分类器的权重通过计算每个子分类器的分类准确率获得.

算法的基本思想是: 首先依据多数类与少数类样本比例, 把多数类样本进行拆分, 拆分时既不增加样本也不删除样本, 同时保持样本的原有空间分布, 然后对拆分后的各个子集, 对每个子集采用 SVM 算法进行训练, 得到多个子分类器, 用每个分类器对训练数据集进行测试, 依据测试准确率来计算每个子分类器的权值, 然后采用 WE 集成分类器方法对各个子分类器进行集成, 最终获得集成后的分类器. 依据委员会投票思想, 奇数个委员更有利于投票结果的产生, 因此本文划分子集的个数为奇数个, 划分子集个数的计算方法后面给出.

2.2 数据拆分

针对此, 本文依据少数类样本数量对多数类样本进行 K-means 聚类, 产生 n 个簇, 从每个簇中取出一个样本构成子集, 每个子集与少数类样本一起构成 m 训练子集, 假设多数类样本 A 的数量为 s, 少数类样本 B 的数量为 n, 划分的子集数为 m, 则 m 计算方式如下:

$$m=\begin{cases}\lceil s/n\rceil & \lceil s/n\rceil/2\ne 0\\ \lceil s/n\rceil+1 & \lceil s/n\rceil/2=0\end{cases} \tag{2}$$

这样设计的 m 值为奇数, 这样便于后面对分类的结果进行判断, 下面给出利用 K-means 算法进行划分数据集的步骤:

算法 1: 拆分数据集

输入: 多数类样本集 A 和少数类样本集 B

输出: 拆分后的子集$T_i, i=1,2,\cdots,m$

Step 1. 对多数类样本进行 K-means 聚类, 产生 n 个簇 $p_1, p_2, \cdots p_n$;

Step 2. 从每个簇 p_i 的样本中随机取出一个样本, 构成子集 T_1;

Step 3. 对剩余样本数$m_i'>0$的簇, 从簇的剩余样本中随机取一个样本; 对于剩余样本数$m_i'<0$的簇, 从簇的初始样本中随机取一个样本, 构成子集$T_j', j=m_i+1,\cdots,m$ 从该簇中取的样本为该簇中随机的一个样本;

Step 4. 重复步骤 3 直到 $i=m$, 对于$m_m'>0$的子集, 则把该簇中剩余的所有样本放到训练集T_m'(最后一个子集)中, 其它簇按照步骤 3;

Step 5. 各个训练子集为$T_i=T_i'+B, i=1,2,\cdots,m$.

2.3 权重计算

权值计算方法是 WE 集成的研究重点, 这里依据每个分类器的测试准确率来计算子分类器的权重, 子

分类器$f_i, i=1,2,\cdots,m$, 对应权重 w_i, 则 w_i 表示为:

$$w_i=\frac{\sum_{j=1}^{s+n}a_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{s+n}a_{ij}} \tag{3}$$

$$a_{ij}=\begin{cases}1, y_j=f_i(x_j)\\ 0, y_j\neq f_i(x_j)\end{cases} \tag{4}$$

其中 $s+n$ 是训练集样本数, $y_j=f_i(x_j)$表示子决策函数 f_i 对样本 x_j 的预测结果与真实结果相同, $f(x_j)\neq f_i(x_j)$ 表示子决策函数 f_i 对样本 x_j 的预测结果与真实结果不相同.

2.4 SCE-SVM 算法

SCE-SVM 算法的流程如图 1 所示, 下面给出算法的详细描述.

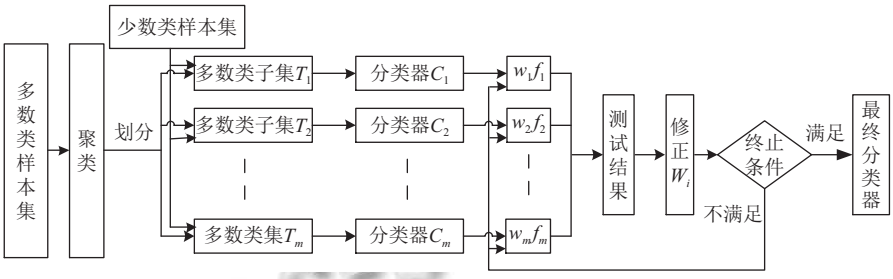

图 1 集成流程

算法 2: SCE-SVM 算法

输入: 不均衡数据集
输入: 决策函数

Step1. 利用算法 1 对数据集进行拆分, 得到对应的训练子集$T_i, i=1,2,\cdots,m$;
Step2. 利用 SVM 算法对各训练子集进行学习, 获得对应的分类器 $C_i, i=1,2,\cdots,m$ 及对应的决策函数 $f_i, i=1,2,\cdots,m$;
Step3. 分别用用决策函数 f_i 对训练数据集进行测试, 依据测试结果利用式(2)计算各子分类器的权值 w_i;
Step4. 依据 w_i 得到集成分类器 $f=\sum_{i=1}^{m}w_if_i$.

3 实验及数据分析

本文中所做实验是在 Matlab 7.11.0 环境下, 结合台湾林智仁老师的 LIBSVM[13], 主机为 Intel Core i7 2.3 GHz, 8 G 内存, 操作系统为 Win7 的 PC 机上完成.

3.1 数据集

该小节的实验在 5 个有代表性的 UCI 数据集上完成, 数据集的基本信息如表 1 所示, 这五组实验数据的类样本间有不同程度的不均衡性, 如表 1 的比例列, 表示多数类样本与少数类样本数量之间的比值, 本实验中多数类样本为正类, 少数类样本为负类. 数据集 letter 是多类数据集, 该实验把其转换为 2 类数据, 数据集 letter 把 A 类作为少数类, 其它 B-Z 类作为多数类.

表 1 实验数据集

序号	数据集	属性	多数类	少数类	比例
1	pima	8	500	268	1.87
2	Phoneme	5	3818	1586	2.4
3	haberman	3	225	81	2.78
4	Balance	4	576	49	11.76
5	letter	16	19211	789	24.39

3.2 实验结果及分析

在不均衡数据的分类方法中, 两类样本的错分代价不同, 更多关注的是少数类的分类准确率, 常用的分类精度不能很好的描述分类器的分类性能, 公认的不均衡数据的评价指标有 F-measure、G-mean 及 ROC 曲线, 用查全率 Recall 和查准率 Precision 来表示, 是公认的不均衡数据的评价标准. ROC 很直观的给出结果, 缺点是不够精确, 因此常采用 ROC 曲线下的面积 AUC 来评价分类器的性能, 该小节通过对比 AUC、F-measure 和 G-mean 来介绍本文算法的有效性. 其中算法 Adaboost、SMOTE、SMOTE-BOOST 的实验结果来自文献[14]. 其中聚类算法是对多数类样本利用 K 均值聚类, 产生与少数类样本数量相同的样本后, 每个簇中心看作一个样本, 与少数类样本合并进行训练.

从 AUC 实验结果来看, 除 Letter 数据集外, 算法 SCE-SVM 都优于其它算法, 可以看到该方法可以用于不均衡数据的分类, 各种算法在 Letter 数据集上的实

验结果差别明显. 由表 3 和表 4 的 F-measure 和 G-mean 的实验结果可以看到, F-measure 和 G-mean 的实验结果比较相似, 算法 SCE-SVM 的实验结果略好于其它算法, 对于 letter 数据集, 各算法的差异较小. 整体看 SVM 算法的实验结果最差, SMOTE 算法较差. 可以看到算法 SCE-SVM 比直接对多数类样本进行聚类的方法要好些. 但是本文算法是基于聚类对多数类数据集进行拆分, 所有数据集具有一定的随机性, 但经过多次实验发现, 对最终的结果几乎没有影响.

表 2 不同算法的 AUC 比较

数据集	聚类算法	SVM	Adaboost	SMOTE	SMOTE Boost	SCE-SVM
pima	0.7139	0.6427	0.7880	0.5571	0.7903	0.8372
Phoneme	0.7313	0.7128	0.9650	0.7714	0.9374	0.9641
Haberman	0.4529	0.4247	0.6411	0.4694	0.6466	0.7038
Balance	0.4176	0.4061	0.6164	0.4528	0.5991	0.6981
letter	0.9813	0.9704	1.0000	0.9851	0.9954	0.9959

表 3 不同算法的 F-measure 比较

数据集	聚类算法	SVM	Adaboost	SMOTE	SMOTE Boost	SCE-SVM
pima	0.6908	0.6349	0.6117	0.5858	0.6416	0.7381
Phoneme	0.7613	0.7531	0.8520	0.6575	0.7965	0.8617
Haberman	0.5000	0.4819	0.3482	0.4877	0.3774	0.5529
Balance	0.2176	0.2097	0	0.1658	0.0011	0.2567
letter	0.7481	0.2212	0.9882	0.9557	0.9664	0.9849

表 4 不同算法的 G-mean 比较

数据集	聚类算法	SVM	Adaboost	SMOTE	SMOTE Boost	SCE-SVM
pima	0.7620	0.6959	0.6946	0.6573	0.7196	0.7501
Phoneme	0.8047	0.7903	0.8901	0.7670	0.8752	0.8992
Haberman	0.6503	0.6389	0.5025	0.6323	0.5362	0.6873
Balance	0.4707	0.4591	0.0012	0.5263	0.0025	0.6694
letter	0.9577	0.9293	0.9892	0.9896	0.9952	0.9941

4 结论

针对实际应用中训练样本不均衡的问题, 对多类数据进行拆分, 构成多个训练子集, 然后对多个分类器进行集成, 得到最终的分类器, 实验数据表明该方法对不均衡数据有较好的分类性能. 如何实现在不均衡数据下的增量学习将是下一阶段的主要工作.

参考文献

1 杜红乐. 基于核空间中 K-近邻的不均衡数据算法. 计算机科学与探索, 2015, 9(7): 869–876.

2 Du HL, Teng SH, Zhang L, *et al*. Support vector machine based on dynamic density equalization. International Conference on Human Centered Computing. Istanbul, Turkey. 2016. 58–69.

3 杜红乐, 张燕. 不均衡数据混合取样分类算法. 燕山大学学报, 2015, 39(2): 158–164.

4 万建武, 杨明, 陈银娟. 代价敏感的半监督 Laplacian 支持向量机. 电子学报, 2012, 40(7): 1410–1415.

5 刘敬, 谷利泽, 钮心忻, 等. 基于单分类支持向量机和主动学习的网络异常检测研究. 通信学报, 2015, 36(11): 136–146.

6 李诒靖, 郭海湘, 李亚楠, 等. 一种基于 Boosting 的集成学习算法在不均衡数据中的分类. 系统工程理论与实践, 2016, 36(1): 189–199.

7 杜红乐, 滕少华, 张燕. 协同标注的直推式支持向量机算法. 小型微型计算机系统, 2016, 37(11): 2443–2447.

8 邢胜, 王熙照, 王晓兰. 基于多类重采样的非平衡数据极速学习机集成学习. 南京大学学报(自然科学), 2016, 52(1): 203–211.

9 林舒杨, 李翠华, 江弋, 等. 不平衡数据的降采样方法研究. 计算机研究与发展, 2011, 48(S2): 47–53.

10 Wang H, Fan W, Yu PS, *et al*. Mining concept-drifting data streams using ensemble classifiers. Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA. 2003. 226–235.

11 Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1): 119–139. [doi: 10.1006/jcss.1997.1504]

12 张春霞, 张讲社. 选择性集成学习算法综述. 计算机学报, 2011, 34(8): 1399–1410.

13 Chang CC, Lin CJ. LIBSVM: A library for support vector machines. (2014). http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

14 李卓然, 张永. 基于集成的非均衡数据分类主动学习算法. 计算机应用与软件, 2012, 29(6): 81–83, 88.