SMOTE 和 Biased-SVM 相结合的不平衡数据分类方法

王和勇[1] 樊泓坤[2] 姚正安[2]

(华南理工大学电子商务学院 广州 510006)[1] (中山大学数学与计算科学学院 广州 510275)[2]

摘 要 针对不平衡数据集的分类问题,本文利用支持向量机推广能力强的优良特性,提出了 SMOTE(Synthetic Minority Over-sampling Technique, SMOTE)和 Biased-SVM(Biased Support Vector Machine, Biased-SVM)相结合的方法。该方法首先对原始数据使用 Biased-SVM 方法,然后对求出的支持向量使用 SMOTE 向上采样方法进行采样,最后再使用 Biased-SVM 方法进行分类。实验结果表明,本文采用的 SMOTE 和 Biased-SVM 相结合的方法可提高不平衡数据集分类精度。

关键词 机器学习,不平衡数据,数据分类,SMOTE, Biased-SVM

Imbalance Data Set Classification Using SMOTE and Biased-SVM

WANG He-yong[1] FAN Hong-kun[2] YAO Zheng-an[2]

(College of E-Business, South China University of Technology, Guangzhou 510006, China)[1]

(College of Mathematics and Computer Science, Sun Yat-Sen University, Guangzhou 510275, China)[2]

Abstract In view of the classification of the imbalance data set, this paper gives the method using SMOTE (Synthetic Minority Over-sampling Technique, SMOTE) and Biased-SVM (Biased Support Vector Machine, Biased-SVM). Firstly, data set is taken into account using Biased-SVM algorithm. Secondly, the sampling to support vector is carried on using the method of SMOTE upward sampling. Finally, the classification is carried on using the method of Biased-SVM. The experimental result indicates that the method given in this paper improves the precision of imbalance data set classification.

Keywords Machine learning, Imbalanced data set, Data classification, SMOTE, Biased-SVM

1 引言

近年来,对不平衡数据集的处理越来越受到专家们的重视,如何有效地从不平衡数据集中挖掘出对人们有用的信息已成为当今机器学习领域研究的热点。不平衡数据集是指某类样本数量明显少于其它类样本的数据集,传统的机器学习分类方法在处理不平衡数据分类问题时会倾向于多数类,而导致少数类的分类精度较低。

国内外学者对不平衡数据集问题的研究已有大量的成果,提出了许多不同的处理方法。其中,有基于数据采样的方法,它们针对数据的预处理,其主要目的是降低数据集的不平衡程度,人为地增加少数类的样本(向上采样,over-sampling)[1],人为地减少多数类的样本(向下采样,under-sampling)[2]等;也有基于分类方法的改进,通过引入某些机制抵消数据不平衡的影响,使其适用于不平衡数据的分类,如 Joshi 等人对 Boosting 算法的改进[3]、Wu Gang 等人对支持向量机方法的改进[4]、Huang Kaizhu 等人提出的 BMPM 算法[5]等。

这些处理方法在不同程度上提高了不平衡数据分类的精度。本文提出了一种改进的不平衡数据分类方法。首先对原始数据使用 Biased-SVM 算法学习找到原始数据的支持向量,然后对求出的支持向量利用 SMOTE 向上采样方法进行采样,最后再使用 Biased-SVM 进行分类,通过试验验证算法的有效性。

2 Biased SVM

在处理不平衡数据分类时,支持向量机(Support Vector Machine, SVM)方法遇到的主要问题是分类面会靠近少数类,从而容易错分少数类。Veropoulos 等人在文献[6]中提出了一种解决这个问题的策略,其主要思想就是赋予错分的正负例不同的惩罚系数 C^+ 和 C^-。这样,目标函数变为

$$\min_{\omega,\omega_0}\frac{1}{2}\|\omega^2\|+C^+\sum_{\{i|y_i=+1\}}\xi_i+C^-\sum_{\{i|y_i=-1\}}\xi_i \tag{1}$$

同样使用拉格朗日乘子法,得到拉格朗日的原始形式:

$$L_p=\frac{1}{2}\omega^T\omega+C^+\sum_{\{i|y_i=+1\}}\xi_i+C^-\sum_{\{i|y_i=-1\}}\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(\omega^T x_i+\omega_0)-1+\xi_i)-\sum_{i=1}^{n}r_i\xi_i \tag{2}$$

其对偶形式与原来的 SVM 完全相同,只是对 α 的约束变为

$$0\leqslant\alpha_i\leqslant C^+,\quad \text{if}\quad y_i=+1$$
$$0\leqslant\alpha_i\leqslant C^-,\quad \text{if}\quad y_i=-1 \tag{3}$$

只要 $\alpha_i=C$,就有 $\xi_i>0$。于是,那些有非零松弛变量的少数类对应的 α_i,要大于有非零松弛变量的多数类对应的 α_i,这样就会将分类面推向多数类。

3 SMOTE 采样方法

SMOTE 采样方法是 Chawla 等人在文献[1]中提出的一种向上采样方法,其主要思想是在相距较近的少数类之间插入"人造"的增加少数类虚拟样本。具体算法如下:对少数类的每一个样本 x,搜索其 k 个最近邻,然后随机选取这 k 个最近邻中的一个设为 $\tilde{x}$,再在 x 与 $\tilde{x}$ 之间进行随机线性插值,构造出新的少数类样本,即新样本

$$x_{new}=x+rand(0,1)\times(\tilde{x}-x) \tag{4}$$

其中 $rand(0,1)$ 表示区间(0,1)的一个随机数。若需要增加更多的少数类虚拟样本,重复以上步骤即可。

图1给出了SMOTE算法的效果，左图为原始样本分布图，右图为使用SMOTE方法增加了两倍虚拟样本后的分布图。可以看出SMOTE方法增加的虚拟样本基本保持了原始样本的分布，但增加的虚拟样本大多分布于原始样本内部，而在边缘附近较少。

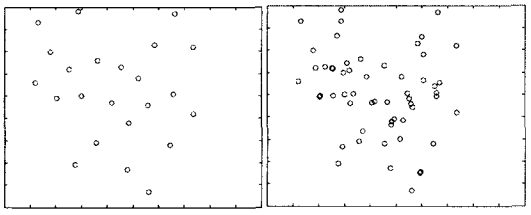

图1 SMOTE方法效果图

4 SMOTE采样和Biased-SVM相结合

本文采用基于支持向量的SMOTE采样方法来处理不平衡数据，然后用Biased-SVM方法进行分类。该方法主要思想是：由于Biased-SVM学习得到的分类面完全由支持向量所决定，因此只需将数据采样的SMOTE方法应用于支持向量，即可减弱不平衡数据集对支持向量集的影响。同时，由于支持向量的数量通常会比较少，因此该方法可以降低系统的响应速度。

根据上述思想，本文的算法流程如图2所示。

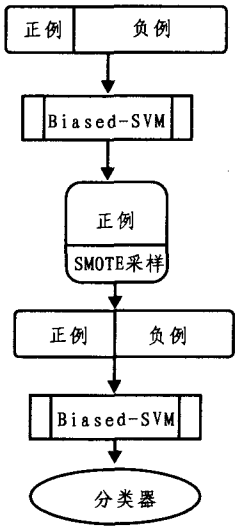

图2 基于支持向量的采样方法流程图

步骤1：用Biased-SVM对原始不平衡的训练集进行学习，记录支持向量。

步骤2：用SMOTE对支持向量进行向上采样，降低多数类和少数类的不平衡程度。

步骤3：再用Biased-SVM进行学习，得到最终的分类器。

5 试验

5.1 试验数据

本文实验所用的数据集是在研究不平衡数据分类时常用的4个公开数据集，它们都从UCI的机器学习数据库[7]中获得，分别是Phoneme，Pima，Satimage，vehicle数据集，其中Phoneme和Pima数据集是两类不平衡问题，Satimage，vehicle数据集原本为多类问题，考虑计算的方便性，本文试验时，对Satimage，vehicle数据集首先转换为两类问题，对Satimage数据集选取类标为"4"的一类数据作为少数类(正类)，将其余的5类合并为多数类(负类)；对vehicle数据集选取类标为"1"的一类作为少数类(正类)，将其余的几类合并为多数类(负类)。这四个数据集的详细描述见表1。

表1 数据集性质

数据集	属性数	总样本数	正例数	负例数	不平衡率	构成
Phoneme	5	5404	1586	3818	2.41	1:0
Pima	8	768	268	500	1.87	1:0
Satimage	36	6435	626	5809	9.28	4:other
Vehicle	18	846	212	634	2.99	1:other

5.2 评价函数

表2是二类问题的混淆矩阵，TP是真实正例的数目，FP是虚假正例的数目，TN是真实负例的数目，FN是虚假负例的数目。

表2 二类问题的混淆矩阵

真实类标 \ 预测类标	正例	负例	Σ
正例	TP	FN	n^+
负例	FP	TN	n^-

一般情况下，人们常用精确率

$$Accuracy=\frac{(TP+TN)}{(n^++n^-)} \tag{5}$$

来评价分类器的性能。但在处理不平衡数据集的分类问题时，这种评价方法显然不合适。举个简单的例子，若数据集极不平衡，仅有5%是正例，那么分类器只需把所有测试样本分为负例，即可获得很高的精确率(95%)，但这样的分类器是没有实际意义的。因此，对于不平衡数据的分类应考虑使用别的评价方法。

本文采用正负例精确度的几何平均，该方法的计算公式为

$$g=\sqrt{acc^+ \cdot acc^-} \tag{6}$$

其中，$acc^+=\frac{TP}{n^+}$为正例的精确率，$acc^-=\frac{TN}{n^-}$为负例的精确率。要使g值大，则需要正负例的精确率都大，且尽量保持两者平衡。如果仅负例的精确率大，而正例的精确率小，那么g值仍会比较小。此外，由式(6)可以看出，g值与acc^+，acc^-的关系是非线性的，acc^+(或acc^-)值的变化对g值的影响依赖于acc^+(或acc^-)值的大小：acc^+(或acc^-)值越小，它引起的g值的变化越大，这就意味着少数类样本错分得越多，错分的代价就越大。

5.3 试验结果

本文采用交叉验证(Cross-validation)的方法[8]，将全部样本随机分为5份，并使每份样本的不平衡率保持与整体不平衡率相等。然后每次取其中4份作为训练集，剩下的1份作为测试集，计算测试集上的g值，把5次g值的平均值作为该算法在整个数据集中的g值。

本文使用标准支持向量机(Standard Support Vector Machine，STSVM)[9]，Biased-SVM[6]，先对正例用SMOTE向上采样，再用Biased-SVM分类(Sos+BSVM)[10]而不是本文提出的对支持向量采用SMOTE向上采样，再用Biased-SVM分类的三种方法与本文提出的方法做比较。表3分别列出了各种方法的g值。

由表3可以看出，在处理不平衡数据分类时，Biased-SVM要明显优于STSVM，尤其是对于属性数和不平衡程度较大的Satimage和Vehicle数据集。Sos+BSVM比单纯使用Biased-SVM的结果高1个百分点，而本文提出的对支持向量采用SMOTE和Biased-SVM相结合的方法(Support Vector SMOTE and Biased-SVM，SVSmt)都比其他方法优越。

表3 各种方法的g值(%)

	STSVM	Biased-SVM	Sos+BSVM	SVSmt
Phoneme	82.698	84.022	84.429	85.193
Pima	71.002	73.276	74.343	75.92
Satimage	75.737	86.764	88.597	89.782
Vehicle	72.158	81.417	82.086	83.232
平均值	75.399	81.370	82.364	83.532

结束语 本文首先介绍了目前对不平衡数据提出的一些处理方法。在此基础上，本文提出了一种基于支持向量的SMOTE向上采样和Biased-SVM相结合的分类方法。通过对UCI四种公开数据集上的对比实验，验证了基于支持向量的SMOTE向上采样和Biased-SVM相结合的分类方法比标准支持向量机方法、Biased-SVM方法和对数据利用SMOTE向上采样和Biased-SVM相结合的方法优越。

参考文献

[1] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 2002,16:321-357

[2] Kubat M, Matwin S. Addressing the Course of Imbalanced Training Sets: One-sided Selection. ICML,1997:179-186

[3] Joshi M, Kumar V, Agarwal R. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements//First IEEE International Conference on Data Mining. 2001: 257-264

[4] Wu G, Chang E. Class-boundary Alignment for Imbalanced Dataset Learning//The Twentieth International Conference on Machine Learning (ICML) Workshop on Learning from Imbalanced Datasets. Washington DC, 2003,8:49-56

[5] Huang Kaizhu, Yang Haiqin, King I, et al. Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004: 558-563

[6] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines //Proceedings of the International Joint Conference on AI. 1999: 55-60

[7] Blake C, Merz C. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/~mlearn/~MLRepository.html. Department of Information and Computer Sciences, University of California, Irvine,1998

[8] Webb A R. 统计模式识别. 王萍,杨培龙,罗颖昕，译. 电子工业出版社, 2004,10:33-37

[9] Vapnik V. The Nature of Statistical Learning Theory. NY: Springer-Verlag, 1995

[10] Akbani R, Kwek S, Japkowicz N. Applying Support Vector Machines to Imbalanced Datasets. ECML 2004, LNAI, 3204: 39-50

(上接第173页)

s^{t+1} \ y^t	Ha	An	Sa	Di
Ha	Su	Fe	Su	Su
Fe	HA	Fe	Sa	Su
Su	Ha	An	Sa	Di

$\{\delta\}$

s^{t+1} \ y^t	Ha	Fe	Su
Fe	Su	Fe	$\{Su, An\}$
Su	Ha	Fe	Su
An	Su	$\{An,Fe\}$	$\{Su,An\}$
SA	Su	$\{Fe,Sa\}$	$\{Sa,Su\}$
Di	Su	An	An

$\{\varepsilon\}$

s^{t+1} \ y^t	Ha	An	Sa	Di
Fe	Su	Fe	Sa	Su
Su	Ha	An	Sa	Di
An	Su	An	An	An
Sa	Su	An	Sa	An
Di	Su	An	Sa	Di

$\{\varphi\}$

图1 情感自动机内部的状态转化

着重对(α)和(β)两图进行解释说明，其他各图依此类推：

(α)中，仅仅当$f(Ha,Ha)=Ha$，其余的情感元中均属于E_0和E_-，而且可以看出$E_-\times E_-\to E_-$，由此可知，当带有负面情感的智能体遭遇到另一带有负面情感的智能体时，它们彼此的情绪都不会有所改观，甚至于交互后，双方情绪会变得更遭，例如，$f(Sa,An)=An$，当处于Sa(忧伤)状态的A_i遇到An(生气)状态的A_j时，就变得十分An(生气)。

(β)中，$f(Fe,Ha)=f(Su,Ha)=Ha$，由此可见当智能体处于E_0状态时，该智能体最容易受周围环境的影响，或者说受另一智能体的影响，从$f(Fe,Fe)=f(Su,Fe)=Fe$和$f(Fe,Su)=f(Su,Su)=Su$上都可以看得出来。特别需要说明的是当$s^{t+1}\in\{Su,An\}$时，$s^{t+1}=Su$和$s^{t+1}=An$的概率均为$\frac{1}{2}$。

结束语 事实证明，情感计算的概念虽诞生不久，却面临着巨大的挑战，原因在于情感的复杂性导致至今为止都没有统一的建模方法，已有的方法大多将情感元素作为一个单一的变量来进行分析和建模，这不符合情感状态复杂多变的特性，本文构建的情感自动机模型不是将某类情感割裂开来，而是把分析个体的主要情感作为研究依据，通过与环境的交互，通过动态Q学习算法，实时地对其自身的情感进行协调和整合，为情感计算提供了方便的研究平台和表现形式，体现了情感计算的发展潜力和应用价值。

参考文献

[1] Michael A. Arbib and Jean-Marc Fellous, Emotions: from brain to robot. Cognitive Sciences, USA ,2004,8(12)

[2] Picard R W. Affective Computing. Cambridge, MA: MIT Press, London, 1997

[3] Hatfield E , Cacioppo J T, Rapson R L. Primitive emotion contagion, Rev. Personal Social Psychol, 1992, 14: 151-177

[4] Hatfield E, Cacioppo J T, Rapson R L. Emotional Contagion. Cambridge University Press, 1994

[5] Andrew, Adamatzky. Affections: automata models of emotional interactions. Applied Mathematics and Computation, 2003, 146: 579-594