非平衡数据集分类研究

吴克寿，曾志强
（厦门理工学院 计算机科学与技术系，福建 厦门 361012）

摘 要：现实世界中存在着非平衡数据集，即数据集中的一类样本数量远大于另一类。而少数类样本的识别通常是人们首要关心的，将少数类样本误分为多数类要比将多数类样本误分为少数类付出更高的代价。传统的机器学习算法可能会产生偏向多数类的结果，因而对于少数类而言，预测的效果会很差。在对目前国内外非平衡数据集研究现状深入分析的基础上，针对非平衡数据集数据复杂度研究和失衡解决方法研究两个方向相对孤立及缺乏系统性的缺陷，提出了一种非平衡数据集整体解决框架，以满足日益迫切的应用需求。

关键词：非平衡数据集；上采样；核学习
中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1673-629X(2011)09-0039-04

Research on Imbalanced Dataset Learning Method

WU Ke-shou, ZENG Zhi-qiang
(Dept. of Computer Science and Technology, Xiamen University of Technology, Xiamen 361012, China)

Abstract: A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world datasets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. It is also the case that the cost of misclassifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. Traditional machine learning algorithms may be biased towards the majority class, thus producing poor predictive accuracy over the minority class. Based on the deep analysis on current research about rare cases classification, proposes a learning framework to address the problem of relative isolation of research between data complexity and solution of imbalanced data, and lack of systematic defects to meet the increasingly urgent applications.

Key words: imbalanced dataset; over-sample; kernel learning

1 非平衡数据集介绍

近年来，非平衡数据集分类问题越来越受到研究者的广泛关注。所谓非平衡数据集，是指同一数据集中某些类的样本数远少于其它类，样本数目少的类称为少数类（这里统称为正类），样本数目多的类称为多数类（这里统称为负类）。由于类别数量上的严重倾斜，传统的分类方法训练所得的结果分类器会对负类样本产生很高的预测准确率，但是对正类样本的预测准确性却很差。即分类器倾向于忽视正类，而正类样本的识别通常是人们首要关心的。例如通过卫星图像检测油井喷发，该数据集中每937张卫星图像中只有41张包含浮油，包含浮油的图像就是正类样本；癌症病人的诊断，将代表正类的癌症病人误诊为健康人，将使其失去及时治疗从而造成难以挽回的后果。类似的非平衡数据集分类问题还包括网络入侵检测[1]、交易欺诈识别[2]、生物数据识别[3]等。事实上，随着信息化发展所带来的知识爆炸，非平衡数据集分类问题广泛存在于工业、农业、科教、国防等各领域，因此，针对性地对其开展研究已成为各领域的迫切需求。

收稿日期：2011-02-17；**修回日期**：2011-05-02
基金项目：国家自然科学基金资助项目（60903203）；福建省教育厅A类科技计划项目（JA08222）
作者简介：吴克寿（1975-），男，湖南长沙人，副教授，博士，研究方向为人工智能、语义网格；曾志强，讲师，博士，研究方向为机器学习、模式识别。

2 国内外研究现状及分析

非平衡数据集分类问题吸引了人工智能、机器学习、数据挖掘领域研究者的广泛关注。计算机领域的著名国际会议AAAI, ICML和SIGKDD分别在2000, 2003和2004年举办了关于非平衡数据集分类问题的专题研讨。在IEEE和ACM主办的著名刊物及会议上所发表的相关文献也逐年快速递增，研究者提出了许多方法来解决非平衡数据集分类问题，但总体来看可归结为两大方向。

2.1 非平衡数据集数据复杂度研究

目前，对非平衡数据集数据复杂度研究主要集中在以下方面：OverLap（两类数据覆盖区域相互重叠），

Rare Instances(正类数据的绝对数量稀少),Small Disjuncts(又称类内不平衡,指同类数据所覆盖区域中存在某些子区域包含较少样本点),噪声。

Garcia[4]通过对多个非平衡数据集的实验分析指出,除类别失衡外,正负两类数据 OverLap 程度越高,则分类器性能越差 。Chen[5] 和 Orriols-Puig[6] 针对 Small Disjuncts 开展研究后指出,不仅类别失衡会影响分类性能,类内失衡同样导致分类性能的降低。Haibo[7]在其文献中阐述了,同正类样本数量相对稀少相比较,Rare Instances 导致的分类器性能下降更为显著,并且更难采用有效方法提高其分类性能。Jason[8] 和 Naeem 等人[9]对噪声进行重点研究后指出:噪声数据对非平衡数据集的分类性能下降起加速作用,并且在很大程度上也影响了改进策略,特别是对抽样学习方法;同负类噪声相比,正类噪声对分类性能下降的影响更为显著。

这部分的研究者大多来自数据挖掘领域。他们的工作揭示了,类别失衡只是影响分类性能的其中一个因素,数据集自身的分布状况如:OverLap,Rare Instances,Small Disjuncts 和噪声等因素都对分类性能产生重要影响,这些因素不仅和类别失衡相互作用从而导致分类器性能低下,而且也影响着非平衡数据集的分类性能提高策略。

2.2 非平衡数据集失衡解决方法研究

这部分研究者大多来自人工智能及机器学习领域,他们专注于数据失衡解决方法研究。

2.2.1 抽样学习方法

抽样学习方法是目前普遍采用解决数据失衡的方法。它分为上抽样(Over-sample,增加样本数量)和下抽样(Under-sample,只抽取部分样本参与训练)。

对于下抽样,最简单的方法就是对负类样本随机抽样以实现类别平衡,然而,这种抽样方式容易导致重要信息的丢失[10]。研究者提出了许多改进方法,如 Yuan[11]首先对负类样本进行聚类,然后采用聚类后的簇质心代替负类样本进行训练,由于聚类后的簇质心在一定程度上反映了原始数据在空间的分布形态,因此优于随机下抽样。刘胥影等[12]将下抽样和集成学习相结合,该方法每次迭代都对负类样本进行随机下抽样,然后训练生成相应的分类器,最后集成所有分类器作为结果分类器,该方法在解决类别失衡的同时避免了重要信息丢失,取得较好效果。Zhang[13]提出一种基于 k 近邻的下抽样方法。该方法选取靠近两类边界的负类样本参与训练,一般认为,靠近边界的样本点包含了较多信息,实验结果证明了该方法的有效性。

最简单的上抽样方法就是采用随机复制正类样本的方式以解决类别失衡。然而这种抽样方法容易导致过学习。Chawla 等人[14]提出 SMOTE(Synthetic Minority Over-sampling Technique)方法,该方法通过在样本与其近邻的连线上随机选取一点作为新样本的方式来增加样本数量。SMOTE 方法所合成样本更为符合原始数据集的数据分布形态,实验取得良好效果,因此受到广泛推崇。研究者后续提出许多以 SMOTE 为基础的改进算法(这里统称为 SMOTE 类型算法),如 Han 提出的针对边界样本的 Borderline-SMOTE[15],Wang 提出的考虑样本近邻情况的 Adaptive Synthetic Sampling SMOTE[16],Jorge 提出了基于距离度量的 SMOTE 类算法[17]。目前,SMOTE 类型算法在抽样学习方法中占主导地位。

2.2.2 代价敏感学习方法

近年来的研究表明,代价敏感学习和非平衡数据集分类两者之间存在紧密联系[18],因此,代价敏感学习的理论和方法可用以解决非平衡数据集分类问题。

Sun[19]将代价敏感学习思路和 Adaboost 算法相结合,将代价项引入 Adaboost 的权值修正策略。Elkan 等人[20]将代价敏感学习引进决策树分类算法,以指导节点分裂及树枝修剪。代价敏感学习也被广泛引入神经网络中,文献[21]中,神经网络的输出根据样本的误分代价而修正,以向正类倾斜,这些研究证实了代价敏感适用于处理非平衡数据集分类问题。

2.2.3 核学习方法

抽样和代价敏感学习方法在前期主要是与决策树分类、神经网络等传统学习算法相结合。近年来,基于结构风险最小化和 VC 维理论的核学习方法已成为人工智能领域的研究热点,它在很大程度上代表了当前机器学习的技术发展水平。支持向量机(Support Vector Machine,SVM)作为核学习方法的代表,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,并且对于非平衡数据集能够提供相对鲁棒的分类性能[22],因此目前被广泛采用与抽样或代价敏感方法相结合来解决非平衡数据集分类问题。和传统学习方法相比,SVM 具有坚实的理论基础及核映射、支持向量等特色,因此极大扩展了抽样和代价敏感学习方法的研究空间。基于此,文中将其作为一种独立的解决类别失衡的方法开展讨论。

由于 SVM 在学习过程中要最小化样本总体分类错误率,因此,当数据集类别失衡程度较高时,SVM 学习完毕所得的最优分类超平面会向正类数据倾斜,从而影响了分类性能。Akbani[23] 和 Wang 等[24] 通过赋予错分的正负类样本不同的惩罚系数来降低分类超平面的偏移度,然而,Imam[25]指出,当正类样本过分稀疏时,采用此种方法会因分类超平面过分拟合正类样本而影响分类效果。Liu[26]通过对负类样本进行下抽样

来降低数据的不平衡率,然而,此类方法可能因支持向量丢失而导致分类性能降低。由于 SMOTE 类型方法在抽样学习算法中占主流地位,因此,许多研究者将其和 SVM 相结合,这些研究都基于支持向量、凸壳、分类间隔等 SVM 特征,如 Han[15] 提出边界点抽样和 SVM 相结合的 Borderline-SMOTE, Vilarino[27] 提出上抽样和 Boost 相结合的 EnSVM(Ensemble of SVM),曾志强等[28]提出的基于核映射的 Kernel-SMOTE, Nguyen[29] 提出基于流形的 Mainfold-SVM,这些算法所采用的抽样方法都是 SMOTE 类型方法,取得较好效果,极大推进了非平衡数据集分类问题研究。

3　存在问题与解决方案

综合以上国内外研究来看,目前非平衡数据集分类问题研究视点较为分散,各研究视点相关性差,缺乏基于相同理论或方法的系统研究,体现在两个方面:

(1)非平衡数据集数据复杂度研究和失衡解决方法研究两个方向相对孤立。以主流的 SMOTE 类型算法为例,虽然目前该类型算法存在许多形式,即多种抽样策略,然而,这些策略都是建立在对数据集分布形态主观假设的基础上,没有对目标数据集进行相应的复杂度分析以指导抽样策略,导致所提出的算法缺乏推广性。

(2)底层的分类算法和上层的抽样及代价敏感等失衡解决方法之间缺乏系统研究。不同的研究者提出不同的失衡解决方法,然后各自在不同的分类算法如决策树、神经网络、贝叶斯分类、SVM 上进行测试,由于各种分类算法基于不同理论,具有不同特征,并且是高度数据依赖性,因此,某种失衡解决方法结合某型分类算法取得较好分类性能,在其它分类算法上不一定能取得同样效果。

基于此,以核学习为基础,以 SMOTE 类抽样算法为主线,以数据集复杂度研究为前提,提出一非平衡数据集分类的整体解决框架,以满足日益迫切的应用需求(选择核学习和 SMOTE 类算法的原因是它们都是各自领域的主流方法,都在处理非平衡数据集分类问题上取得较好效果)。

该学习框架如图 1 所示。首先开展数据集复杂度研究,获得数据集复杂度分析算法和知识库,接着开展噪声处理研究,实现噪声处理算法。在知识库的基础上进行抽样算法研究,实现抽样算法,并进行抽样与代价敏感学习算法的比较研究,建立理论模型。由于本学习框架是基于核方法,故采用 SVM 作为学习机,由于 SVM 具有较高的训练和分类复杂度,因此有必要开展 SVM 快速训练与分类算法研究,获得相应算法,最后集成所有算法形成非平衡数据集分类解决框架,开展应用。

数据集复杂度研究方案如图 2 所示,首先采集或构造大量实验数据集,建立衡量数据复杂度及分类性能的各个量化指标,采用 SVM 对实验数据进行分类学习。在此实验数据基础上,运用频繁关联规则,支持向量回归机等构建各指标与分类性能之间的因果关系,建立相应知识库,以指导后续抽样算法研究。

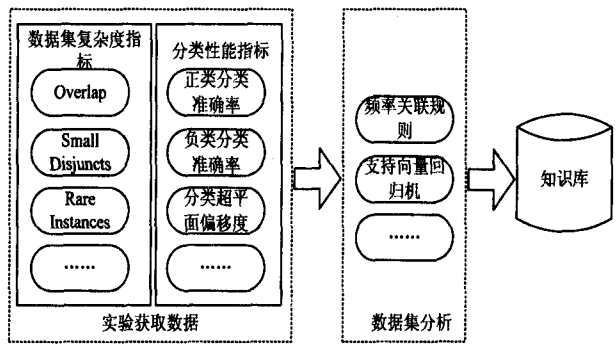

图 1　非平衡数据集分类学习框架

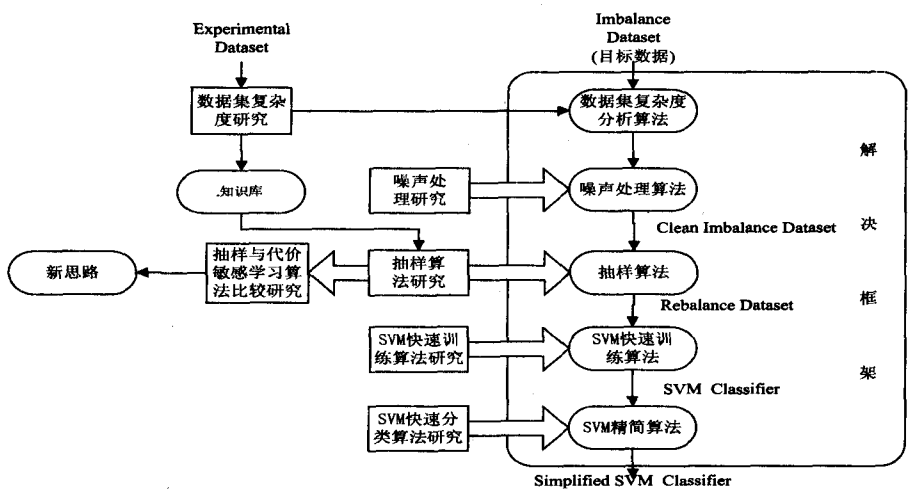

图 2　数据集复杂度研究

4　结束语

文中在对目前国内外非平衡数据集研究现状深入分析的基础上,揭示了当前研究工作存在的局限性,并提出了一种基于核方法的非平衡数据集学习框架,为进一步从理论和实际的结合上研究非平衡数据集分类问题开拓新思路。

参考文献:

[1] Lazarevic A, Ertoz L, Ozgur A, et al. Evaluation of Outlier De-

tection Schemes for Detecting Network Intrusions [C]//In Third SIAM International Conference on Data Mining. [s. l.]: [s. n.],2003.

[2] Fawcett T, Provost F. Combining Data Mining and Machine Learning for Effective User Profiling[C]//In Simoudis, Han, Fayyad. The Second International Conference on Knowledge Discovery and Data Mining. [s. l.]: AAAI Press, 1996: 8-13.

[3] Muggleton S H, Bryant C H, Srinivasan A. Measuring Performance When Positives Are Rare: Relative Advantage versus Predictive Accuracy —A Biological Case-study[C]//In European Conference on Machine Learning. [s. l.]: [s. n.], 2000: 300-312.

[4] Garcia V, Mollineda R A, S'anchez J S. On the k-NN performance in a challenging scenario of imbalance and overlapping[J]. Pattern Analysis Applications, 2008, 11(3-4): 269-280.

[5] Chen Mu-chen, Chen Long-sheng, Hsu Chun-chin. An information granulation based data mining approach for classifying imbalanced data[J]. Information Sciences, 2008, 178: 3214-3227.

[6] Orriols-Puig A, Bernad'o-Mansilla E. Evolutionary Rule-based Systems for Imbalanced Data Sets[J]. Soft Computing, 2009, 13(13): 213-225.

[7] He Haibo, Garcia E A. Learning from Imbalanced Data[J]. IEEE Transaction on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[8] Van Hulse J, Khoshgoftaar T. Knowledge Discovery from Imbalanced and Noisy Data[J]. Data & Knowledge Engineering, 2009, 68(12): 1513-1542.

[9] Seliya N, Khoshgoftaar T M, Van Hulse J. A Study on the Relationships of Classifier Performance Metrics[C]//Proceedings of 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2009). Newark, New Jersey, USA: [s. n.], 2009.

[10] 李 鹏, 汪晓龙, 刘远超, 等. 一种基于混合策略的失衡数据集分类方法[J]. 电子学报, 2007, 35(11): 2161-2165.

[11] Yuan J, Li J, Zhang B. Learning concepts from large scale imbalanced data sets using support cluster machines[C]//Proc. of the ACM Int'l Conf. on Multimedia. [s. l.]: [s. n.], 2006: 441-450.

[12] Liu X L, Wu J, Zhou Z H. Exploratory Under Sampling for Class Imbalance Learning[C]//Proc. Int'l Conf. Data Mining (ICDM 2006). [s. l.]: [s. n.], 2006: 965-969.

[13] Zhang J, Mani I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction[C]// Proc. Int'l Conf. Machine Learning (ICML '2003), Workshop Learning from Imbalanced Data Sets. [s. l.]: [s. n.], 2003.

[14] Chawla N V, Bowyer K W, Hall L O, et al. Synthetic minority over-sampling technique[J]. J. Artif. Intell. Res, 2002, 16: 321-357.

[15] Han H, Wang Y H, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [C]//Proc. Int'l Conf. Intelligent Computing. [s. l.]: [s. n.], 2005: 878-887.

[16] Wang B X, Japkowicz N. Imbalanced Data Set Learning with Synthetic Samples[C]//Proc. IRIS Machine Learning Workshop. [s. l.]: [s. n.], 2004.

[17] Calleja J, Fuentes O. A Distance-Based Over-Sampling Method for Learning from Imbalanced Data Sets[C]//In International Florida Artificial Intelligence Research Society (FLAIRS) Conference. [s. l.]: [s. n.], 2007: 634-635.

[18] Chawla N V, Japkowicz N, Kolcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6.

[19] Sun Y, Kamel M S, Wong A K C, et al. Cost-Sensitive Boosting for Classification of Imbalanced Data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[20] Elkan C. The Foundations of Cost-Sensitive Learning[C]// Proc. Int'l Joint Conf. Artificial Intelligence. [s. l.]: [s. n.], 2001: 973-978.

[21] Kukar M Z, Kononenko I. Cost-Sensitive Learning with Neural Networks[C]//Proc. European Conf. Artificial Intelligence. [s. l.]: [s. n.], 1998: 445-449.

[22] Japkowicz N, Stephen S. The Class Imbalance Problem: A Systematic Study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449.

[23] Akbani R, Kwek S, Japkowicz N. Applying Support Vector Machines to Imbalanced Data Sets[J]. Lecture Notes in Computer Science, 2004, 3201: 39-50.

[24] Wang B X, Japkowicz N. Boosting Support Vector Machines for Imbalanced Data Sets[J]. Lecture Notes in Artificial Intelligence, 2008, 4994: 38-47.

[25] Imam T, Ting K M, Kamruzzaman J. z-SVM: An SVM for improved classification of imbalanced data[C]//Australian Joint Conference on AI. Hobart, Australia: Springer, 2006: 264-273.

[26] Liu Y, An A, Huang X. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles [C]//Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006). Singapore: Springer Press, 2006: 107-118.

[27] Vilarino F, Spyridonos P, Radeva R, et al. Experiments with SVM and Stratified Sampling with an Imbalanced Problem: Detection of Intestinal Contractions[J]. Lecture Notes in Computer Science, 2005, 3687: 783-791.

[28] 曾志强, 吴 群, 廖备水, 等. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报, 2009, 37(11): 2489-2495.

[29] Nguyen. Learning from Categorical and Numerical Imbalanced Data[D]. Japan: Japan Advanced Institute of Science and Technology, 2006.