基于集成金字塔模型的单分类方法

薛 文 谢永红 马延辉 杨炳儒
(北京科技大学信息工程学院 北京 100083)

摘 要 针对单分类提出了一种集成金字塔模型(Single-Class Classification Integrated Pyramid Model,SCCIPM),由综合获取层、辅助判定层、核心分类层和结果优化层4个独立协同的层面组成。其中综合获取层由改进的KNN以及优化的1-DNF两个分类方法组成,主要用来获取反例,其结果均提交到辅助判定层投票后得到可靠反例;核心分类层通过多次迭代更新可靠反例,每次迭代建立一个分类器;结果优化层根据核心分类层所建立的不同分类器优化选择最终分类器。仿真实验表明,当正例占整个样本的50%以下时,SCCIPM较其它方法优势明显,在解决单分类上具有良好的分类性能。

关键词 集成金字塔模型,单分类,数据挖掘
中图法分类号 TP18 **文献标识码** A

Single-Class Classification Approach Based on Integrated Pyramid Model

XUE Wen XIE Yong-hong Ma Yan-hui YANG Bing-ru
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract A Integrated Pyramid Model is proposed for Single-Class Classification(Single-Class Classification Integrated Pyramid Model, SCCIPM). This model is composed of four independent collaborative layers by Comprehensive obtain layer, Assistant judgment layer, Kernel classification layer and Optimization layer. Comprehensive obtain layer formed by improved KNN and optimized 1-DNF two classification methods, mainly for obtaining negative examples, results are submitted to Assistant judgment layer voting reliable negative examples. Kernel classification layer update reliable negative examples by several iteration, each iteration establish a classifier. Optimization layer according to Kernel Classification layer of different classifiers, optimally select the final classifier. The simulation experiment result indicates that when positive examples are 50% of total samples below, SCCIPM shows obvious advantages over other methods in solving the Single-Class Classification and has good classification performance.

Keywords Integrated pyramid model, Single-class classification, Data mining

1 引言

数据挖掘的分类方法中,传统的二分类问题需要两类样本训练分类器,首先是要给出正例和反例集合,然后算法利用它们构造分类器,再用这个训练好的分类器对测试样本进行分类处理。

在实际分类器的构造过程中,训练样本的正例相对比较容易获得,而反例的获得却存在很多问题,如反例获取代价较高:对于卫星网络故障诊断,要获取故障样本所付出的代价太高,不可能为了获取故障样本而特意让卫星系统出现故障;还有就是反例样本数目几乎无穷:比如入侵检测中攻击数量和种类层出不穷,而且有的攻击产生变种,根本就不清楚该拿哪些攻击样本来训练入侵检测器。这些问题都使得获得反例困难重重,而且获取的反例也不一定准确。由于反例集合获取(复杂性或代价)难,很多情况下只能通过获取正例集合建立分类器,因此形成了单分类问题[1]。

虽然反例的获得确实很难,但大量未标记样例却可以容易收集。关于利用未标记样例集合改善学习性能,提高分类器分类精度,近年来研究工作者提出了一些切实有效的方法,包括B. Liu提出的NB(Naive Bayesian)[2],Yu,Han和Chang提出的PEBL (Positive Example Based Learning) [3]。Fung等人提出了PNLH(Positive examples and Negative examples Labeling Heuristics)[4],证明了未标记样例集合确实能够更好地改善学习性能,提高分类效果[5]。但这些方法仍然存在稳定性差,当正例占整个样本非常少时,不能有效地进行分类。针对这些问题,本文提出一种新的多层递阶分类系统模型——集成金字塔模型来解决单分类问题。该模型主要由综合获取层、辅助判定层、核心分类层和结果优化层4个独立协同层面组成。综合获取层融合了数据挖掘中改进的KNN以及优化的1-DNF方法以获取反例,辅助判定层投票获得可靠反例,综合获取层通过支持向量机方法建立多个分类器,核心分类层进行最终的优化选择。本文模型使得分类器的构造不

到稿日期:2010-07-19 返修日期:2010-10-15 本文受国家自然科学基金(60675030,60875029)资助。
薛 文(1985-),女,硕士生,主要研究方向为数据挖掘,E-mail:xuewen.ustb@gmail.com;谢永红(1970-),副教授,硕士生导师,主要研究方向为数据库和数据挖掘、人工智能;马延辉(1985-),硕士生,主要研究方向为数据挖掘;魏明伟(1987-),硕士生,主要研究方向为数据挖掘。

需要人工收集反例集合，仅由已经标记的正例集合和一个全局的未标记样例集合来构造，其中未标记样例集合中的每一个样例都是随机选取的，而且数量通常远远大于标记的正例数目。结果表明，集成金字塔模型能够很好地对单分类问题进行分类。利用层次间的智能接口将各种方法有机融合，从而充分发挥各层模型与方法的优势，不仅减少了对标识训练集中类别数据大小要求，还利用大量廉价易得的未标记数据集辅助分类，使整体效果达到最优。

2 集成金字塔模型

对于单分类这类非传统分类问题，一般单一的分类模型或多个单一模型简单组合构成的分类模型均难以获得令人满意的分类结果。集成金字塔采取了逐步求精、分层递阶结构，各个层次各有侧重、功能独立且通过智能接口紧密衔接，因此集成金字塔模型具有相对传统分类方法更高的分类准确度，同时具有更好的普适性。其模型架构如图 1 所示。

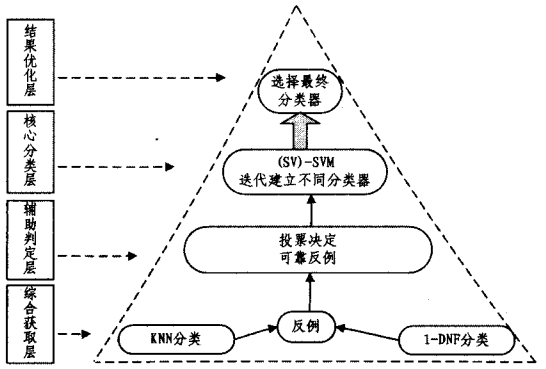

图 1 集成金子塔模型

图 1 中集成金字塔模型由综合获取层、辅助判定层、核心分类层以及结果优化层构成，各层功能独立、紧密协同。值得注意的是，集成金字塔模型与传统的机器学习（如仅利用标识正例集合的监督学习方法[6]以及仅利用未标记集合的无监督学习[7]）不同，同时考虑了正例和未标记集合的信息，而且利用了数据挖掘不同的方法知识进行学习，因此具有更高的准确度。

在综合获取层，由两种方法把一部分反例得到初步的预测，辅助判定层通过上层两种方法的结果共同进行决策，获取可靠反例[8]。核心判定层的功能通过正例和获取的可靠反例进行迭代，构建分类器。在每次迭代中，对未标记集合循序渐近地学习和标记，从而完成对分类器参数的调整，更新分类器。这样，在每一轮的迭代过程中，随着未标记集合规模的减小标记样本数量的增大，学习器不断适应当前环境下的分类，直到未标记集合中无样本可以标记，最终完成对未标记集合的分类。结果优化层在核心分类层中的每一步迭代都可建立一个不同的分类器。由于数据噪声等因素，可能对一些样本错误标记，这样多次迭代后最后的一个并不一定是最好的分类器，通过方法进一步判断，选择最终分类器。为了描述简单，以下用 P 代表标记的正例集合，U 代表未标记集合。

2.1 综合获取层

该方法构造初始训练集合，需要一定数量的反例。本层综合了改进的 KNN [9]和优化的 1-DNF [10]方法进行反例获取，是整个模型的基础层。

2.1.1 改进的 KNN 方法

对于反例的获取，采用对于 U 中的某一样例 U_d 在 P 中找到与此最近的一个正例 P_d，根据 k 值，在 P 中找到与 P_d 最近的 k 个邻居，求出 k 个邻居之间的平均距离 P_{mean}，然后计算 P_d 与 P_{mean} 之间的距离。如果此距离与 U_d 和 P_d 之间的距离相除小于某一设定阈值 δ，则判断 U_d 属于反例。但在计算 k 个邻居之间的平均距离时，用传统的 KNN 方法会造成 k 个邻居对最后平均距离影响都一样，使得方法对于 k 的选取很敏感。降低 k 影响的一种有效途径就是根据每个邻居的距离不同对其作用加权。权重的产生采用高斯函数，该函数在距离为 0 时所得的权重值为 1，并且权重值会随着距离的增加而减少，权重值始终都不会跌至 0。实现加权 KNN 与普通的 KNN 最重要的区别在于，它并不是对这些元素简单地求平均，它求的是加权平均。加权平均的结果是通过将每一项的值乘以对应权重，然后将所得结果累加得到。待求出总和以后，再将其除以所有权重之和。

伪代码如下：

Input：P，U

Output：反例 N_{knn}

Algorithm：

1. 设置一个值域 δ 和 k 的值；
2. for(each U_d in U)
3. compute U_d 与其最近的正例 P_d 之间的欧式距离 $D_1=d^2_{U_d,P_d}=||U_d-P_d||_2^2$；
4. 在 P 中，compute P_d 与其 k 个最近的正例 $\{P_{1NN},\cdots,P_{kNN}\}$ 之间的欧式距离 $d^2_{P_d,P_{iNN}}=||P_d-P_{iNN}||_2^2, i=1,2,\cdots,k$ 和相应权重 $w_{P_d,P_{iNN}}=\exp(-d^2_{P_d,P_{iNN}}/2\mu^2) i=1,2,\cdots,k$（其中 μ 为核参数，其决定了权值变化趋势的快慢）；
5. compute 这个 k 个邻居的加权平均

$$P_{mean}=\frac{\sum_{i=1}^{k} w_{P_d,P_{iNN}} \cdot P_{iNN}}{\sum_{i=1}^{k} w_{P_d,P_{iNN}}};$$

6. computeP_d 与 P_{mean} 的欧式距离 $D_2=||P_d-P_{mean}||_2^2$；
7. If D2/D1$<\delta$
   $U_d\in N_{knn}$；
   else
   $U_d\notin N_{knn}$；
8. return 反例 N_{knn}。

2.1.2 改进的 1-DNF 方法

本文方法对于反例获取，采用根据特征在 P 的频率大于在 U 的频率来构造正例特征集合 PF。U 中的某一样例的特征包含 PF 中正例特征的出现数目小于阈值 ρ，则认为此样例是反例。但是仅仅考虑了一个特征在 P 和 U 中出现频率的差异，从频率差异的角度来定义所谓的正例特征是不全面的。比如，一个特征在 P 中出现的频率为 1%，而在 U 中出现的频率为 0.99%，这样的特征明显不能够很好地代表 P，因此需要考虑其本身在 P 中出现的绝对频率。将绝对频率的阈值设定为 γ%，规定同时满足下述两个条件的特征才可加入到 PF 中：一是在 P 中出现的频率大于其在 U 集合中出现的频率；二是在 P 中出现的频率大于所设定的阈值。

伪代码如下：

Input：P，U

Output：反例 N_{1-dnf}

Algorithm：

1. 确定训练样本集合中的特征集合 $w_i(i=1,2,\cdots,n)$，其中 $w_i\in U\cap P$；
2. for($i=1;i<=n;i++$)

3. if($freq(w_i,P)/|P|>freq(w_i,U)/|U|\&\&freq(w_i,P)/|P|>\gamma\%$)
4. $PF=PF\cap w_i$;
5. for(each U_d in U)
6. if($\forall w_j\in PF\&\&freq(w_j,U_d)<\rho$)
7. $N_{1-dnf}=N_{1-dnf}\cup\{U_d\}$;U=U−{$U_d$};
8. return 反例 N_{1-dnf}。

其中,$|P|$为 P 的数量,$|U|$为 U 的数量,$freq(w_i,P)$为特征 w_i 在 P 中出现的次数,$freq(w_i,U)$为特征 w_i 在 U 中出现的次数。

2.2 辅助判定层

根据综合获取层 KNN 与 1-DNF 两种方法提取的反例的结果,采用某种投票的原则决策最终的可靠反例。本文采用的投票策略是当两分类器都把样例 U_d 划分到反例时,认为 U_d 属于可靠反例。

Input:N_{knn},N_{1-dnf}
Output:可靠反例 RN
Algorithm:
1. if ($U_d\in N_{knn}\&\&U_d\in N_{1-dnf}$)
2. $U_d\in RN$;
3. return 可靠反例 RN。

2.3 核心分类层

核心分类层中,根据正例集合和获取的可靠反例建立初始的 SVM 分类器。对 U 进行分类,把获取的反例添加到可靠反例中,迭代重新建立分类器。然而,如 U 很大,训练的时间就会很长。简单地缩减 U 的样本,虽然会减少低运行的时间,但同样会损失准确性,因为 U 直接影响到训练结果的质量。这里提出一种方法,能够很明显地减少训练的时间,同时能保持住准确性。基本思想主要是在每一次迭代后去掉对分类决策影响很小或者根本没有影响的训练数据,对于 SVM 建立分类器只有位于间隔区边缘的数据才是确定分界线位置所必需的,可以去掉其余所有的数据。而分界线依然还是位于相同的位置,将位于这条边界线附近的数据称作支持向量[11]。因此在每次迭代的时候,去掉远离间隔区边界的数据,这样虽然 U 的数据量减少了,但却不会影响算法的准确性。本文采用(SV)-SVM((Support vector)-Support vector machine))方法进行迭代,生成不同的分类器。

伪代码如下:

Input:P,U,RN
Output:分类器 SVM_i
Algorithm:
1. 对 P 中每条记录赋予标记 positive label;
2. 对 RN 中每条记录赋予标记 negative label;
3. 使用 P 和 RN 训练新的 SVM 分类器 SVM_i;
4. 用分类器 SVM_i 对 $Q(U-RN)$进行分类;
5. 对 Q 中分类为反例的样本,组成集合 W;
6. $Q=Q-W$;$RN=W$;(而不采用 $RN=RN\cup W$);
7. goto(3)直到 W=NULL;
8. return SVM_i。

2.4 结果优化层

在核心分类层中,每一次迭代都建立了不同的 SVM 分类器,但最后一个并不一定是最好的分类器。因为在现实中,训练数据集中可能包含噪声样例,而 SVM[12]对噪声数据敏感。如果在每一次迭代中从 $Q(U-RN)$中错误地获取了一些正例放到 RN 中,那么最后生成的 SVM 分类器(SVM_{last})就会不准确,决定采用哪个分类器,根据 SVM_{last}对 P 进行分类。如果有 $\xi\%$(文献[5]为 8%)的 P 被分为反例,那么就说明 SVM_{last}不准确,则用第一个 SVM 分类器(SVM_1),否则用 SVM_{last}。

伪代码如下:

1. 用 SVM_{last}对 P 进行分类;
2. if > $\xi\%$的 P 被分为反例 then
   用 SVM_1作为最终分类器;
   else
   用 SVM_{last}作为最终分类器;
3. return 最终分类器。

3 实验

3.1 评价指标

正例和反例的标号分别是 positive 和 negative,TP 和 TN 分别是正确分类的正例和反例的样本数量,FN 和 FP 分别是误分类的正例和反例的样本数量。表 1 是两类数据集的混合矩阵。

表 1 两类数据集的混合矩阵

	分类类别为 positive	分类类别为 negative
实际类别 positive	TP	FN
实际类别 negative	FP	TN

根据混合矩阵,可以定义以下参数:

$$accuracy_{pos}=\frac{TP}{TP+FP}$$

$$accuracy_{neg}=\frac{TN}{Tn+FN}$$

$$G\text{-}means=\sqrt{\frac{TP\cdot TN}{(TP+FN)\cdot(TN+FP)}}$$

其中,$accuracy_{pos}$和 $accuracy_{neg}$分别是正例和反例样本的准确率,$G\text{-}means$ 综合考虑了正例和反例的分类准确率;$G\text{-}means$ 指标越大,表明分类效果越好。

3.2 实验环境与数据集

为了验证模型的正确性和有效性,本文在 UCI 的多个数据集上进行测试。实验硬件环境为联想 2.97GHz/2G 个人计算机,实验所涉及的主要软件环境为 Microsoft windowsXP,SQL server 2005,visual C++ 6。

实验中涉及到的数据集见表 2。其中的 data set 列是数据集的名称,attribute 列记录了数据集中记录的描述属性个数,class 列为数据的分类属性个数,Number of Instances 列记录了数据集中记录的个数。

表 2 数据集

data set	attribute	class	Number of Instances
breast	10	2	699
heart	13	2	270
auto	25	2	205
sick	29	2	2800
Sea	18	2	1350

实验中随机选取数据集中每个类别的 70%作为训练集合,剩余的 30%用作测试集合。在训练集合中,随机选取 $\lambda\%$的正例样本作为实验中的 P 集,剩余$(1-\lambda\%)$的正例样本和其他类别的样本组成实验所需的 U 集。

3.3 实验结果与分析

为了验证 SCCIPM 的分类性能,用 SCCIPM 与 NB,PE-

BL,PNLH 进行测试,结果如图 2 所示。图 2(a)—图 2(c)依次是上述 5 个数据集中在不同 λ%情况下各方法对于所有测试类别的 $accuracy_{neg}$,$accuracy_{pos}$ 和 *G-means* 值的平均变化趋势。

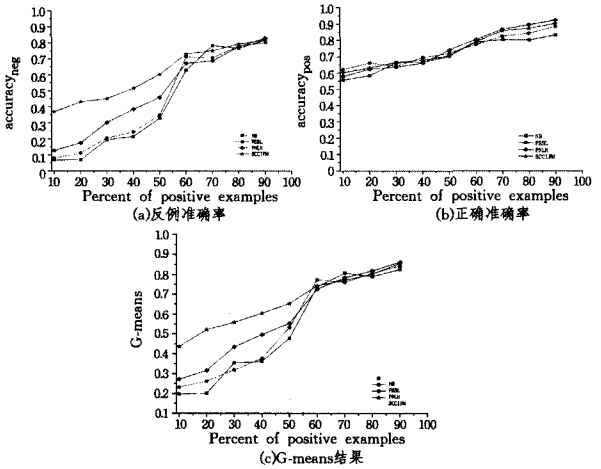

图 2　正例规模对精度的影响

由此可见,随着正例占整个样本的百分比不断增加,所采用的 3 个评价指标均在不同程度地增加。通过增加数据集中正例样本的数量,改变训练数据集所给出的信息,试图有效提高反例样本分类精度,由此提高分类器的整体性能。

通过图 2(a)可以看到,在反例的分类精度方面,SCCIPM 在正例样本占整个样本的百分比越小时,较其他方法越显示出较强的分类能力,从而达到了较理想的分类效果。这一发现对解决"提高分类器精度"这一单分类的瓶颈问题有着重要的意义。

图 2(b)表明增加正例样本数量对正例样本分类精度影响不是特别显著,但仍是增加的趋势,而且一直保持在较高值上。无论正例占整个样本的比例有多大,正例的分类精度都高于反例的分类精度。在正例样本分类精度方面,SCCIPM 仍然在正例少的情况下分类性能仍优于其它分类器。

进而从图 2(c)得出,SCCIPM 对样本的整体分类性能最优;且与其它分类器相比,不同数据集中整体的样本分布状况、样本规模、特征维度数以及正例占整个样本的百分比等因素对 SCCIPM 影响最小。当正例占整个样本的 50%以下时,优于其它分类器;在正例占整个样本的 50%以上时,优势不是很明显,和其它方法取得了相似的效果,但 NB 和 PEBL 都不大稳定。这也表明了 SCCIPM 分类性能更稳定,鲁棒性更强。这充分验证了本文提出的 SCCIPM 模型在单分类问题中具有强的普适性,从而对从根本上解决单分类问题具有十分重要的意义。

结束语　为了更好地发现单分类问题的本质,解决"提高分类器精度"这一单分类中的瓶颈问题,本文提出了基于单分类环境下集成金字塔模型。该模型特点是不需要人工收集反例集合,只使用标记的正例集合和一个全局的未标记样例集合来构造分类器,存在着容易收集训练样例集合的优越性,有着广阔的应用前景。对比实验结果表明,SCCIPM 在单分类上具有较强的分类效果。但模型还需要在更多的真实数据集上进行相关实验,以充分检验其健壮性和有效性。

参 考 文 献

[1] Yu H. Single-class classification with mapping convergence[J]. Machine Learning,2005,61(1-3):49-69

[2] Liu B,Dai Y,Li X L,et al. Building Text Classifiers Using Positive and Unlabeled Examples[C]// ICDM-03. Melbourne,Florida,November 2003:19-22

[3] Yu H, Han J, Chang K C-C. PEBL: Positive Example Based Learning for Web Page Classification Using SVM[C]// Proc. Ninth Int'l Conf. Knowledge Discovery and Data Mining. 2003

[4] Fung G P C,Lu H J. Text Classification Without Negative Examples Revisit[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(1):6-20

[5] Zhang Bang-zuo,Zuo Wan-li. Learning from Positive and Unlabeled Examples: A Survey[C]// International Symposiums on Information Processing and 2008 International Pacific Workshop on Web Mining and Web-based Application. Moscow,Russia, May 23-25,2008:640-644

[6] 李和平,胡占义,吴毅红,等. 基于半监督学习的行为建模和异常检测[J]. 软件学报,2004,18(3):527-537

[7] Figueiredo M A T,Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Trans Patt Anal Mach Intell,2002,24(3):381-396

[8] Zhang B Z,Zuo W L. A Novel Reliable Negative Method Based on Clustering for Learning from Positive and Unlabeled Examples[C]//Proceedings of the Asia Information Retrieval Symposium 2008(AIRS 2008). Harbin,China,January 2008:393-400

[9] Baily T,Jain A K. A note on distance-weighted k-nearest neighbor rules[J]. IEEE Trans. Syst. Man Cybern. ,1978, SMC-8(4):311-313

[10] 于海龙. 面向 PU 问题的文本分类的研究与实现[D]. 长春:吉林大学,2005

[11] Cortes C, Vapnik V. Support-Vector-Networks [J]. Machine Learning,1995,20(3):273-297

[12] Han J-W,Kamber M. Data mining: concepts and technique(2nd ed)[M]. Morgan Kanfmann,2007

(上接第 182 页)

[5] Kang K,Kim S,Lee J,et al. FORM: A Feature-oriented Reuse Method with Domain-Specific Reference Architectures[J]. Annals of Software Engineering,1998,5:143-168

[6] Griss M L, Favaro J, D'Alessandro M. Integrating Feature Modeling with the RSEB[C]// ICSR98. Victoria, BC, IEEE, June 1998:36-45

[7] Eriksson M,Borstler J,Borg K. The PLUSS Approach-Domain Modeling with Features,Use Cased and Use Case Realizations [C]// Proceedings of the 4th International Conference on Software Product Lines(SPLC'OS). LNCS 3714,Springer-Verlag, 2005:33-44

[8] 张伟,梅宏. 一种面向特征的领域模型及其建模过程[J]. 软件学报,2003,14(8):1345-1356

[9] van Harmelen F A G. Web ontology language:OWL[C]//Staab S,Studer R,eds. Handbook on Ontologies in Information Systems. Springer-Verlag,2003:67-92

[10] Martin D,Burstein M,et al. OWL-S:Semantic markup for Web services[OL]. http://www.daml.org/services/owl-s/1.1/overview/,2003

[11] 赵毅,胡丹. FODA:一种面向特征的领域分析方法[J]. 重庆工学报学报:自然科学版,2004,18(5):45-47