基于半监督技术的多分类器融合策略研究

蔡 晰[1,2],郭躬德[1,2],黄添强[1,2]
CAI Xi[1,2],GUO Gong-de[1,2],HUANG Tian-qiang[1,2]

1.福建师范大学 数学与计算机科学学院,福州 350007
2.福建师范大学 网络安全与密码技术重点实验室,福州 350007
1.School of Mathematics and Computer Science,Fujian Normal University,Fuzhou 350007,China
2.Key Lab of Network Security and Cryptography,Fujian Normal University,Fuzhou 350007,China
E-mail:caixisea@163.com

CAI Xi,GUO Gong-de,HUANG Tian-qiang.Multiple classifiers fusion method based on semi-supervised learning.Computer Engineering and Applications,2009,45(25):218-221.

Abstract: This paper proposes a novel strategy for multi-classifier classification.The method takes maximal error correcting ability as a criterion of choosing classifiers.To improve the accuracy of multi-classifier classification,a semi-supervised co-training technology is employed which makes use of the complementarity of each single classifier and maximizes the judging ability of the arbiter as well.The experimental results show the mothod is practical and effective on real toxicity dataset.
Key words: multi-classifier classification;co-training;arbiter;semi-supervised learning

摘 要:提出一种新颖的多分类器构造方法,它以最大纠错能力作为分类器选择标准。实现时,采用半监督协同训练技术,充分利用单分类器的互补性,同时最大化仲裁器的仲裁能力,以提高多分类器系统的分类精度。在毒性数据集上的实验结果表明了方法的可行性和有效性。
关键词:多分类器;协同训练;仲裁器;半监督学习
DOI:10.3778/j.issn.1002-8331.2009.25.067 **文章编号**:1002-8331(2009)25-0218-04 **文献标识码**:A **中图分类号**:TP18

1 引言

在现实生活中,化工产品在批量生产前需要经过非常严格的毒性检测,保证其对环境(自然,人,动物)的无害性。传统的做法往往在动物身上进行测试,然而,有些化工产品对人、动物或自然的破坏,是一个循序渐进的过程,要得到可靠的实验数据,需要很长的实验周期,同时需要耗费大量的财力[1]与物力,并进行长时间的跟踪和观察。有些社会急需的化工产品,如药品,因此无法进入市场,导致许多患者病情耽搁甚至死亡。所以,使用科学方法减短实验周期具有重大的现实意义。

化工产品的毒性强弱与其内部的化学结构密切相关,例如,一些药品的毒性强弱与苯环,光能团的个数相关。因此,人们希望从已知毒性的化工产品中找出这些属性与毒性的内在联系,因此分类技术很自然地被引入到该领域中。从目前已知的化工产品属性和毒性的关系中建立模型,对新的化工产品毒性进行预测,预测结果对决策部门制订有关法规,出台有关政策具有重要的参考价值。因此,提高化工产品毒性预测的精度显得至关重要[2]。

单分类器由于其本身的方法和结构特点,在对数据进行分类预测时存在很大的局限性。多分类器由于综合了各个单分类器的优点可以较好地解决这个问题,从而获得比单分类器更高的分类精度[3]。

由于传统的毒性测试方法需要耗费大量的财力、物力,已知毒性的化工产品的数量有限,仅仅利用已知毒性的化工产品难以构造良好的分类模型,因此需要额外的信息来共同构造良好的模型。在现实中存在大量的未知毒性的化工产品,它们的属性(化学结构等)是容易得到的,这些属性隐藏着有用信息,研究充分利用这些化工产品的有用信息来构造更优的分类模型,以得到更精确的分类结果。

提出一种基于半监督技术的多分类器融合策略,该策略在基础分类器的选择上,采用最大纠错能力作为分类器选择标准,使单分类器的互补性最大,并且最大化仲裁器的仲裁能力;在半监督学习阶段,为了充分利用单分类器的互补性,采用半监督协同训练技术,从而达到比采用自监督技术更高的分类精度,并在一些实际应用的毒性数据集上进行验证。

2 相关背景

2.1 多分类器中基础分类器的选择和融合

2.1.1 基础分类器的选择

在构造多分类器的过程中,需要对基础分类器进行选择和

基金项目:福建省自然科学基金(the Natural Science Foundation of Fujian Province of China under Grant No.2007J0016)。
作者简介:蔡晰(1985-),男,研究生,主要研究领域为:数据挖掘与人工智能;郭躬德(1965-),男,教授,博士生导师,主要研究领域为:人工智能、机器学习和数据挖掘技术及其应用;黄添强(1971-),男,副教授,主要研究领域为:数据挖掘技术与地理信息系统。
收稿日期:2008-05-15 **修回日期**:2008-07-28

组合,穷举是最直接也是效果最好的办法,但缺点是时间复杂度大,目前大量采用的方法是根据相似性度量选取差异性较大的分类器进行组合,另外遗传算法,纠错能力优先等算法也常用在基础分类器的选择上。

最简单的选择单分类器的方法就是把所有的可能都列举出来,然后依次进行实验选择出效果最好的作为最佳的分类器组合。在分类器数量比较小的情况下,使用选择测试方法是非常有效的,但是在分类器数量很大的情况下穷举是不现实的。普遍的做法是固定分类器的数量,然后根据一定的标准与原则进行选择,这些原则中最常用的就是"差异性最大原则"。

Zhou[4]等人提出使用遗传算法来进行分类器的选择的GASEN算法,具体的思想是通过重复抽样产生不同的基础分类器,对每一个分类器指派一个初始权重,然后通过遗传算法来调整权重,使权重的分布接近全局最优,最后保留权重大于阈值的分类器。Geng和Zhou[5]还提出了基于本征空间的选择SEME方法。该方法的核心是先通过初步筛选选取 n 个识别率较高的候选分类器,然后,根据能否将已经选中的分类器分错的例子分对作为进一步挑选的依据,也就是挑选那些纠正错误能力较强的分类器。初始时集合中只有一个分类精度最高的分类器,然后逐步选入纠错能力最强的分类器,直到达到规定的分类器数量。Parikh[6]等通过调整训练集的权重来判断各个分类器对于全局的贡献,训练集中的每一个例子,如果在上一轮被分正确,它的权重减少,如果被分错,它的权重不变,这样每轮能对那些分错的例子分对的基础分类器权重增加,以此来选择单分类器。

2.1.2　基础分类器的融合

在多分类器的融合阶段,有很多常用的方法,如投票法,动态选择方法,基于相似度的方法,贝叶斯融合方法和基于模糊集的方法等。投票法就是根据一定的算法算出各个分类器在最后投票时的权重分布,最后综合所有分类器的投票结果作为最后的输出;动态选择法就是根据不同的实例寻找适合的分类器对它进行分类。Giacinto和Roli[7]提出了一种动态选择的方法,它首先用单分类器的分类结果作为融合阶段的属性,对于一个新的实例,算出与它距离比较近的一些点(用融合阶段的属性作为度量),用对这些点分类效果最好的分类器对新的点进行分类。类似的方法还有方[8]提出的动态分类器选择算法DSg和动态投票算法DVSg,它根据单分类器的预测结果来判断要预测的实例的邻居,再由此选出分类效果最好的分类器或者确定各个分类器的权重。刘等[9]提出了通过考察重叠区域的可能性来自适应调整分类器的权值;Huang[10]提出了行为知识空间法,该方法不要求分类器具有独立性。具体的说是根据单分类器对训练集的分类结果,形成由 ***BKS*** 向量组成的行为知识空间,每一个 ***BKS*** 向量对应一种单分类器决策组合,向量的内容包括符合该决策组合的训练样本个数,符合该决策组合的训练样本中占主导地位的类别以及每一类的训练样本总数,对于要预测的实例,根据单分类器的决策结果找到对应的 ***BKS*** 向量,其中占主导地位的类别就作为该实例的预测值。

2.2　半监督学习

在训练集数量很少的情况下,很难建立一个好的模型进行预测。在实际应用中,数据集中往往包含大量的未标签的实例,人们尝试各种各样的方法,把未标签的例子所包含的信息利用起来,加入到训练集中以获得更好的分类模型,这就是半监督学习。一般认为半监督学习开始于Shahshahani和Landgrebe 1994年的工作[11]。半监督学习的一个重要组成部分是自监督学习。自监督学习是指分类器通过原始的训练集把未标签的实例进行标签,然后挑选一些加入到本身的训练集中,使训练集的数量增大,从而建立更好的模型。Zhou[12]应用这种思想在已标签数据非常少的情况下取得了很好的分类效果。另一种重要的半监督学习方法是协同训练,最早的协同训练算法是Blum和Mitchell[13]1998提出的。算法要求问题有充分冗余的两个视图,例如网页上的信息和该网页的链接,在两个视图上分别训练出一个分类器,然后挑选若干置信度高的实例加入到对方的训练集中,从而提高分类精度,最后选取置信度大的分类器对新的实例进行分类。由于充分冗余的两个视图这个条件很难满足,Nigam和Ghani证明在属性很大的情况下可以随机地把属性集分为两个视图,也能取得很好的效果[14]。Goldman和Zhou[15]提出使用不同的决策树算法在同一属性集上训练,然后两个分类器进行联合训练,互相挑选置信度高的例子给对方,最后由置信度最高的分类器来分类,避免了对属性集的要求。Zhou和Li于2005年提出了tri-training算法,该算法不要求属性集类型不同,也不要求基础分类器类型不同,它使用重复抽样来形成3个不同的训练集,通过相同的算法形成3个基础分类器,在协同训练阶段,如果两个分类器对未标签的例子判断相同就把该例子加入到第3个分类器的训练集中,最后通过投票来进行集成[16]。除此之外,协同训练技术也用在半监督回归中[17]。

在选取基础分类器时对最大纠错原则进行改变,在第三个分类器的选择上,不考虑对前两个分类器都分错的实例进行纠错,而仅仅考虑对前两个分类器意见不一的实例的纠错能力,这样在限定分类器数量为三的条件下达到全局最优。另外在半监督阶段采用协同训练的方法利用未标签实例的信息,最后用投票加仲裁的融合方法确定多分类器的最后输出。

3　基于半监督技术的多分类器融合算法

3.1　问题描述

每一种化工产品可表示为$\{x_1,x_2,x_3,\cdots,x_{n,y}\}$其中 $x_1,x_2,\cdots,x_n$ 表示药品的 n 个属性,y 表示类标签,类标签有3种{1,2,3}分别代表强毒性,中等毒性,弱毒性。已知毒性的化工产品称为已标记的实例,未知毒性的化工产品称为未知实例。C_1,C_2,C_3 表示3个不同的分类器,其中 C_1,C_2 作为基础分类器,C_3 作为仲裁分类器。$\boldsymbol{X}$ 表示属性向量$(x_1,x_2,x_3,\cdots,x_n)$,$y_i=C_i(X)$表示第 i 个分类器对实例的预测结果。

3.2　前期的数据处理

正如前言所说,未标签的实例往往是廉价的且容易得到,在实际应用中未标签实例数量往往很大,如文本分类。因此挑选合适的未标签实例加到训练集中是非常耗时的。有的文章采用可重复抽样[5]的方法来减少未标签实例的数量以降低时间复杂度,但是没有约束的抽样对抽样结果是没有保证的,即可能筛选掉真正有用的实例而留下对分类没有帮助的未标签实例。为了解决这个矛盾,基于DBSCAN聚类方法的思想,挑选一些密度较大的点构成未标签实例的候选集合,因为密度较大的点在被分类时的置信度较高,对分类有帮助的可能性较大,选择这样的例子作为未标签集合,可最大限度地保留对分类有帮助

的未标签实例。

3.3 分类器的选择和联合训练

自监督是一种重要的半监督方法,其主要思想是在训练集数量不足以形成很好的分类模型的情况下,从未标签的实例中选取一些置信度高的实例,加入到自己的训练集中,使自己的训练集数量增大,从而建立更好的分类模型,提高自己的分类精度。在一些应用中自监督取得了很好的效果,但该文采用联合训练而不使用自监督技术,主要原因是在基础分类器分类精度不高的条件下,两个分类器互补性强更为重要。自监督是从未标签的实例中选取置信度高的实例加入到自己的训练集中,这种做法相当于从自己的优势区域中挑选出一些未标签的实例来加强自己本来分类精度高的区域,而对自己原本分类能力弱的区域提高有限,综合起来分类精度只能得到有限的提高。

在分类器的选择上,采用最大纠错能力的原则,即选择分类精度最高的单分类器作为第一个分类器 $C1$,然后选择对 $C1$ 分错的实例分类精度最高的分类器作为第二个分类器 $C2$(对 $C1$ 的纠错能力最强),这种选择策略使得两个分类器具有很强的互补性。互补性强是指在 $C1$ 分类精度较高的区域 $C2$ 的分类精度未必高,而在 $C1$ 分类精度较低的区域 $C2$ 的分类精度应较高。这里采用联合训练的方法来达到这一目的,让 $C1$,$C2$ 互相弥补各自的不足,即最大限度的利用它们之间的互补性达到全局的分类精度的提高。

为了最大限度的利用两个分类器的互补性,需要让它们互相学习对方的优点(但不学习缺点)。$C1$ 从未标签实例中挑选一些它擅长区域的实例,对这些实例进行标签。由于是 $C1$ 的擅长区域,这部分数据的置信度较高。而这个区域恰恰是 $C2$ 的弱势区域,把 $C1$ 挑选出来的实例加入到 $C2$ 的训练集中就能在很大程度上提高 $C2$ 在这个区域的分类能力,从而提高 $C2$ 整体的分类精度。这样两个分类器的互补性得到充分利用,精度都得到了提高。联合训练图例见图 1 所示。

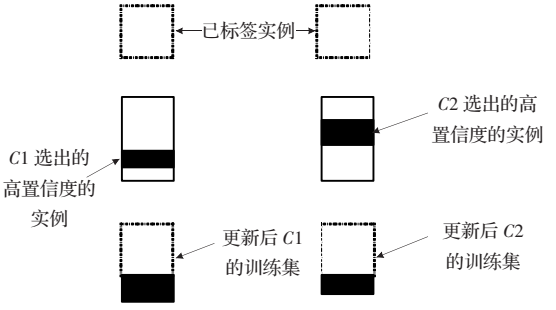

图 1 联合训练

3.4 仲裁器

仲裁器的作用是在 $C1$,$C2$ 的分类结果不一致的情况下由 $C3$ 来决定测试实例属于哪一类。因此对 $C3$ 的要求就是对 $C1$,$C2$ 分类结果不一致的实例分类精度尽量高,也是最大纠错能力原则。采用如下的方法:将训练集按 70%,30%分为新训练集 D 和新测试集 T,新测试集 T 中挑选出 $C1$,$C2$ 分类结果不同的实例组成测试集 $T1$,用各个单分类器对 $T1$ 进行分类,分类精度最高的一个单分类器被选为仲裁分类器。这个分类器对在 $C1$,$C2$ 确定的前提下具有比全局分类精度最高的单分类器更高的纠错能力,从而保证了多分类器的整体最优,见图 2。

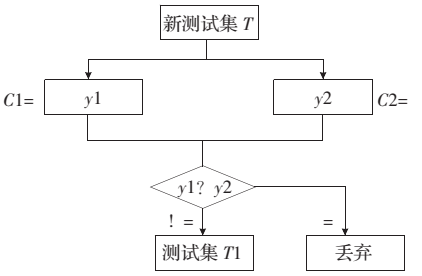

图 2 仲裁器

3.5 测试阶段

在测试阶段时,对于要测试的实例,首先由联合训练后的 $C1$,$C2$ 对新的实例进行分类,如果 $C1$,$C2$ 对这个实例的分类结果相同,就把它分到这个类,否则由仲裁分类器 $C3$ 决定它属于哪个类,见如图 3。

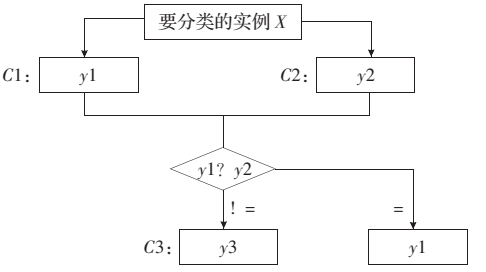

图 3 测试阶段

3.6 算法

输入:已标签实例集合 L,未标签实例集合 U,分类器 $C1$,$C2$,$C3$;

输出:分类结果。

训练阶段:

步骤 1 $C1$,$C2$ 的训练集初始化为 L;

步骤 2 $C1$ 对 U 中的每一个实例进行预测,并取出预测置信度大于阈值 h 的实例集合 $L1$,$C2$ 对 U 中的每一个实例进行预测,并取出预测置信度大于阈值 h 的实例集合 $L2$;

步骤 3 $C1$ 的训练集更新为 $L\cup L2$,$C2$ 的训练集更新为 $L\cup L1$;

测试阶段:

对于要测试的实例 X

if ($C1(X)$==$C2(X)$)

then $y=y1=y2$;

else $y=C3(X)=y3$;

4 实验

4.1 数据集

在对 5 种毒性测试数据集进行实验以验证该文提出的多分类器算法的性能,Bee,Daphnia,Dietary,Oral,Trout,数据来自欧共体 DEMETRA 项目[18]。每个数据集都有一些已标签的实例

表 1 实验中使用的数据集的一些基本信息

项目	数据集				
	Bee	Daphnia	Dietary	Oral	Trout
训练集数量	71	182	83	80	181
未标签实例	289	132	273	277	115
测试集数量	31	77	35	34	95

表 2　实验结果

数据集	Bee	Daphnia	Dietary	Oral	Trout	Average
单分类器 *C*1 的分类精度/(%)	41.94	66.23	65.71	76.47	57.89	61.65
采用的算法	KNN4	KNN6	KNN6	KNN3	KNN4	
单分类器 *C*2 的分类精度/(%)	45.16	57.00	57.14	73.53	56.84	57.93
采用的算法	NaiveBayes	NaiveBayes	NaiveBayes	NaiveBayes	NaiveBayes	
单分类器 *C*3 的分类精度/(%)	35.48	65.00	48.57	58.82	60.00	53.57
采用的算法	LWL	SMO	J48	LWL	KNN1	
平等投票的分类精度/(%)	35.48	62.34	57.00	73.53	61.00	57.87
*C*1 自监督的分类精度/(%)	**48.39**	66.00	**68.00**	76.47	56.84	63.14
*C*2 自监督的分类精度/(%)	35.48	54.54	60.00	73.53	58.95	56.50
自监督的多分类器的分类精度/(%)	38.70	66.23	65.71	76.47	62.00	61.82
该文提出的方法的分类精度/(%)	**48.39**	**68.83**	65.71	**79.41**	**65.26**	**65.52**

和一些未标签的实例,由于已标签的实例数量不足,单纯使用已标签实例构造模型进行预测难以取得令人满意的效果,因此采用半监督技术将未标签实例的信息加以利用,构建更好的分类模型,实验中将已标签的实例随机地分为训练集(约 70%)和测试集(约 30%),具体情况如表 1 所示。

原始的数据集共有 226 种属性, 经过特征选择取出重要的 10 种,加上分类属性一共 11 维属性。类标号统一设置为 3 种。由于未标签的事例数量不多,所以没有使用聚类抽样方法进行预处理。在实验中,使用分类精度来衡量分类算法的好坏的标准,实验中置信度阈值取 0.995。

4.2　实验结果

对每个数据集分别采用了单分类器, 单分类器半监督,平等投票,自监督多分类器方法 与提出的方法进行实验,具体结果列在表 2 中。

单分类器 *C*1,*C*2,*C*3 分别表示 3 个单分类器的分类精度,“平等投票”表示 3 个单分类器进行平等投票的分类精度,随后的五项中“*C*1 自监督”表示单分类器 *C*1 采用自监督方法后的分类精度,“*C*2 自监督”表示单分类器 *C*2 采用自监督方法后的分类精度,“自监督的多分类器”表示采用自监督方法的多分类器的分类精度。该文提出的方法表示采用联合训练方法的多分类器的分类精度。

4.3　结果分析

单分类器 *C*1*C*2 采用半监督方法后的分类精度没有得到很大提高,单分类器是否采用半监督对分类精度没有本质的影响,未采用半监督的平等投票法与单分类器相比,分类精度也没有明显地提高,将半监督方法与多分类器结合的自监督多分类器算法与联合训练多分类器方法都取得了令人满意的效果,总体表现要优于其他的分类算法。对比最后的分类精度,由于联合训练的多分类器方法更好地利用了单分类器的互补性,因此分类结果略优于自监督的多分类器方法。

5　总结

改造了最大纠错原则的分类器选择方法,采用联合训练的半监督技术,应用委员会决策加仲裁的结构,形成一种多分类器方法。在毒性数据集上与平等投票法,自监督仲裁算法进行比较实验,取得了较好的实验结果。

实验结果说明该方法在基础分类器分类精度不高的条件下比较适用,在基础分类器分类精度较高的情况下的效果有待进一步研究。此外,在基础分类器判断时仅采用两个分类器的联合训练,采取多个基础分类器进行联合训练是一个可能的改进方法。在半监督的过程中,采用置信度准则选取未标签的实例后,可以再对这些实例进行一次过滤,以最大限度地排除噪声。在联合训练中,采用一定的限制进行迭代也是一种可能的改进方法。

致谢:感谢福建省自然科学基金(2007J0016),福建省教育厅 A 类科技项目(JA07051)以及英国工程物理科学研究基金(GR/T02508/01)的赞助。

参考文献:

[1] Xu L,Wu Y,Hu C,et al.A QSAR of the toxicity of aminobenzenes and their structures[J].Science in China:Series B,2000,43(2):130-136.

[2] 李丹,李国正,陆文聪.用于药物活性预报的 Co-Training 方法[J].计算机科学,2006,33(12):159-161.

[3] Neagu D,Guo G D.A data-driven approach for improved effective classification in predictive toxicology[C]//Proc of IEEE-ICCC(IEEE International Conference on Computational Cybernetics),Tallinn,Estonia,2006.

[4] Zhou Z H,Wu J,Tang W.Ensembling neural networks:Many could be better than all[J].Artificial Intelligence,2002,137(1-2):239-263.

[5] Geng X,Zhou Z H.Selective ensemble of multiple eigenspaces for face recognition[R].AI Lab,Computer Science & Technology Department,Nanjing University,Nanjing,China,2003-08.

[6] Parikh D,Kim M T,Oagaro J,et al.Combining classifiers for multisensor data fusion[C]//2004 IEEE International Conference on Systems,Man and Cybemetica,2004,2(10-13):1232-1237.

[7] Giacinto G,Roli F.Dynamic classifier selection based on multiple classifier behavior[J].Pattern Recognition,2001,34(9):1879-1881.

[8] 方敏.集成学习的多分类器动态融合方法研究[J].系统工程与电子技术,2006,28(11):1760-1769.

[9] 刘汝杰,李华胜,袁保宗.基于自适应权值的多分类器融合方法[J].北方交通大学学报,2001,25(2):14-17.

[10] 王伟.多分类器日文假名识别研究[D].哈尔滨:哈尔滨工业大学,2004.

[11] Shahshahani B,Landgrebe D.The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J].IEEE Transactions of Geoscience and Remote Sensing,1994,32(5):1087-1095.

(下转 230 页)