学校编码：10384　　分类号__密级___
学号：23220060153381　　UDC___

厦门大学

博 士 学 位 论 文

带置信度分类器的研究与应用

Research and Application of Classifier with Confidence

王 华 珍

指导教师姓名：林 成 德 教 授
专 业 名 称：控制理论与控制工程
论文提交日期：2009 年 4 月
论文答辩时间：2009 年 5 月
学位授予日期：2009 年 月

答辩委员会主席：_____
评 阅 人：_____

2009 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

高风险领域的分类问题对模式分类算法提出以下三个挑战：

1）能否设计一种分类器模型，使得它的输出结果能够附带置信度；

2）预测输出的置信度是有效的，应该使得算法的准确率能够被置信度所控制。

3）算法应能够独立地对每个测试数据提供相应的置信度评估，也就是说，能够根据指定的置信度产生相应的预测结果。

针对上述三个挑战，我们引入了基于转导推理和随机性检验的置信预测方法来解决这些问题。近年新发展起来的一致性预测器（CP）是这种方法的典型代表。但是，CP 在实践中的实用性较差，主要是其固有的运算效率低下、以及对样本奇异函数的设计缺乏指导性准则。我们的工作主要是改进了 CP 的理论模型，提出了混合压缩一致性预测器（HCCP）的算法框架及其实现技术，使其更适合于实际应用。

HCCP在预测性能与计算效率间取得了一个较好的折衷，它在处理大数据集学习问题时，在保持算法的预测效率的同时大大提高了CP的运算效率。HCCP的运作特点是将用于学习的样本序列划分成两个部分，并采用两阶段混合压缩：第一阶段先将前一部分序列样本压缩成一个模型，并以知识的形式保存；第二阶段再将上述知识传递给后续的检验样本序列用于置信预测。在算法实现技术方面，HCCP采用有监督的度量学习方法来实现有效信息在两个子序列（训练样本序列和检验样本序列）中的传递。并分别通过有监督核学习方法（HCCP-KerNN）和随机森林技术（HCCP-RF）实现了度量学习和样本奇异函数的设计。我们还从实验角度展现了HCCP-RF算法在田纳西－伊斯曼化工过程（TEP）这样的流程工业大系统的在线故障置信检测中的适用性和有效性。

针对小数据集的学习问题，我们也提出了一种无划分的 HCCP-RF 算法，它取消了对学习样本集的划分，更加适用于处理小样本数据。通过对慢性胃炎中医诊断数据集的实验，验证了该算法的有效性。

最后，对本文的工作进行了总结，并对今后的研究工作提出了展望。

关键词：分类问题;置信预测;一致性预测器

Abstract

There are three challenges to the researchers on the classification in the high-risk areas:

1）Can we develop a classification algorithm that outputs predictions coupled with confidence level?

2）Are these confidences for the predictions really valid, i.e., could the accuracy rate be guaranteed by the confidence level?

3）Could the algorithm give a prediction with a confidence level tailored for each individual instance, in other words, could it provide a prediction corresponding to the confidence level predefined?

Faced to these challenges, we have introduced a method which uses the transductive inference and the randomness test of i.i.d. sequences to develop our solution. The recently emerged Conformal Predictor (CP) is an alternative solution which can output prediction with valid confidence. However, there are still certain disadvantages in the framework of CP, such as the inherent computational costliness and the lack of guidance for the design of the example nonconformity measure. We have focused on the improvement and the enhancement of CP, and have then proposed a new Hybrid-Compression Conformal Predictor (HCCP) which performs better in practice.

HCCP aims to obtain a good balance between the predictive performance and the computational efficiency. It can maintain a relatively high predictive performance while improving greatly the computational efficiency in dealing with large data sets. HCCP divides the whole training examples into two subsets (called as the training set and the validation set, respectively) and executes the predicting process in two stages. Firstly, it abstracts a compression model *M* based on the training set; secondly, it designates, for each example in the validation set, the new features which are generated by *M* and would then be applied by the classical CP algorithm to output the prediction with confidence level. We have proposed a method based on the supervised

metric learning to transfer the useful information from the first stage to the second stage. In detail, we have incorporated the adaptive kernel-based distance metric learning method (as in HCCP-KerNN) and the random forest algorithm (as in HCCP-RF), respectively, to realize the supervised metric learning and the example nonconformity measure. The application is simulated on the standard large data set as Tennessee Eastman Process (TEP). The applicability and effectiveness of the proposed HCCP-RF algorithm are illustrated on this online fault detection of large-scale industrial process.

To deal with the problem of small-sample classification, we have also put forward the non-partition HCCP-RF algorithm, which disclaims the partition of the whole learning set of examples. The application is simulated on the traditional Chinese chronic gastritis data set, which is a typical small-sample problem. The informative as well as effective predictions of the non-partition HCCP-RF algorithm have been shown in the experiment.

Finally, the summary of our work and the future research are presented.

Keywords: classification problem, prediction with confidence, conformal predictor

目录

Contents

第1章　引言

1.1 基于机器学习的分类问题

1.1.1　分类问题描述

随着信息技术和存储技术的快速发展，政府、商业、企业等各行各业出现了越来越多的复杂非线性高维数据。如何根据用户的特定需求从海量数据中发现有用的知识或者构造从经验中学习的机器，用于对未来数据进行预测成为一个十分迫切的富有挑战性的研究课题。机器学习(Machine Learning, ML)致力于让计算机模仿人类从实例中学习的能力进行数据分析和建模（估计某系统隐藏的、复杂的输入输出关系），使它（这种关系）能够对未知输出做出尽可能准确的预测[1]。

模式分类是实际应用中普遍存在的问题，也是机器学习领域的基础研究之一。现实生活中存在大量的分类问题，如：机械故障诊断、医学诊断、语音识别、人脸识别、信用评估、文本分类、网络入侵检测、图像识别等。分类的作用和根本目的在于面对某一未知类别的具体事物时，能按照已知的信息将其正确的归于某一类。将某一研究对象正确归入某一类的方法即分类方法。在机器学习领域，模式分类致力于从有限观察发现观测数据中暗含的各种关系，具体说就是从实际问题的一个有限的子集（样本集）出发，探求问题的内在规律（建立模型），从而对未知数据做出正确判断（分类）。这就是机器学习领域的分类问题（classification）。

分类问题用数学语言可以简单描述如下：实际问题的具体对象一般有很多属性，可用高维向量$x=(x^1,x^2,\dots,x^d)^T\epsilon X$ 表示，其中数据x的上标$1,2,\dots,d$是向量的维数序号，也就是对象的属性序号；$X=\mathbb{R}^d$表示 d 维实数空间。假设实际问题对应有限c个可选类别，用标识变量$y\epsilon Y=\{1,2,\dots,c\}$表示对象的类别，其中$Y$叫做类别空间。对象及其类别构成数据对$z=(x,y)\epsilon Z$，$Z$叫做样本空间，$z_i,i=1,2,\dots,n,\dots$是样本空间中的样本（点），其下标“$i$”表示样本编号。模式分类定义为：

定义 1.1 (*分类问题* ***F***) 根据给定的训练样本集$Z^{(n)}=\{z_1,z_2,\dots,z_n\}$,其中$Z^{(n)}$表示含有$n$个样本的样本集,$z_i\epsilon Z, i=1,\dots,n$表示样本，产生一个分类器$\varphi:\mathbb{R}^d\longrightarrow\{1,2,\dots,c\}$，使得它对新的待测试数据$x_{n+1},x_{n+2},\dots,x_{n+k}$的实际类别值产生相应的预测值$\hat{y}_{n+1},\hat{y}_{n+2},\dots,\hat{y}_{n+k}$。我们要求该分类器要能对整个样本空间可能的分布有一个尽可能小的期望判别误差。

从定义 1.1 看，模式分类技术模仿了人类的逻辑推理过程。逻辑推理遵循的一般性途径大概分为两种方式：即归纳推理和演绎推理，它们两者的相互关系见图 1.1

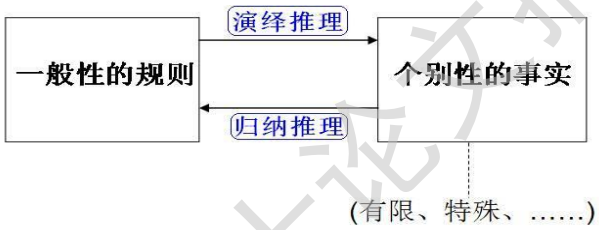

图 1.1 逻辑推理示意图

由上图可知，模式分类的学习过程是归纳推理，它从个别性的事实（一般是有限个的，具有特殊代表性）出发归纳出一般性的规则，而分类器的预测过程是演绎推理，它从一般性的规则推理出个别性的事实。

1.1.2 分类问题研究进展

随着 1968 年 K.Popper 用不可证伪的概念提出了关于归纳问题的理论[2]，统计学被看作是归纳推理的一个数学模型，因此统计学也成为模式分类的重要理论基础之一。在此基础上已经出现了多种以经典统计理论为工具刻画的模式分类方法，如：贝叶斯决策、K-近邻、线性判别分析、决策树等等[3-5]。这些经典的统计机器学习方法在机器学习问题中起着基础性的作用。然而统计分类器偏离了统计推理的本质模式（它应该是根据观测数据寻求感兴趣的分类器（函数）），而变成具有 R.Fisher 参数估计特点的‘模型辨识’方法。Fisher 统计是基于参数估计的统计推理，他把从给定数据估计函数这个问题（分类问题、回归问题和密度估计问题）表达为特定（参数化）模型的参数估计问题，并提出了估计所有模型未知