**软化纠错输出编码多分类算法 郭宏洲 指导教师 刘昆宏 厦门大学**

**2029320381452802095400**

**本 科 毕 业 论 文 （设 计）**

**（主修专业）**

**软化纠错输出编码多分类算法**

**Soft Coding ECOC Multiclass Classification Algorithm**

姓 名： 郭宏洲

学 号： 24320142202424

学　 院： 软件学院

专 业： 软件工程

年 级： 2014级

指导教师： 刘昆宏 副教授

**二〇一八年五月**

**厦门大学本科学位论文诚信承诺书**

本人呈交的学位论文是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，并符合相关法律规范及《厦门大学本科毕业论文（设计）规范》。

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）。

本人承诺辅修专业毕业论文（设计）（如有）的内容与主修专业不存在相同与相近情况。

学生声明（签名）：

年 月 日

致 谢

摘 要

通过对大量数据的特征以及对应数据的类别进行分析，以此得到的模型或公式可以用于预测未知数据的类别，而这类问题被称为分类问题。分类问题从简单的二分类问题到复杂的多分类问题在诸多领域中均有广泛的应用，并且根据其原理，已经开发了许多不同的方法以解决这些问题。目前分类问题主要研究的内容是多分类问题，而面对多分类问题，以多个简单分类器为基础，以多种方式将它们结合的集成学习方法是目前较为主流和有效的方式。纠错输出编码(Error Correcting Output Codes，简称ECOC）是其中一种集成学习方法。和决策树类似，ECOC亦是通过将多分类问题分解成多个二分类问题，再对每个二分类问题构造一个适合的基础二分类器中用以解决，再以这些二分类问题的结果为依据，确定多分类问题的结果。

在多分类问题中，偏标签分类问题是一类特殊的问题，它具体的表现在每个训练样本与一组候选标签相关联，而其中只有一个是对应于样本的真实标签。近年来，进行对偏标签数据学习和分类的需求自然地出现在许多现实世界的应用中。本文则提出了一种以适用偏标签数据的K最近邻分类器为先验信息并结合以数据复杂度为依据的特征空间选择的算法来提高ECOC对偏标签数据的分类能力。

本文将改进过后的偏标签ECOC算法与现有的偏标签ECOC算法在一些偏标签真实数据集上的分类准确率进行对比，同时还对这些真实数据集的数据分布进行了分析，从而证明了改进过后的偏标签ECOC算法效果确实更加理想。

关键词：纠错输出编码；先验信息；数据复杂度

# Abstract

**Key words:** ECOC; Soft Coding; Ensemble Learning

目 录

[第一章 绪论 1](#_Toc514264496)

[1.1 引言 1](#_Toc514264497)

[1.2 论文组织结构 3](#_Toc514264498)

[第二章 相关概念介绍 5](#_Toc514264499)

[2.1 多分类问题 5](#_Toc514264500)

[2.2 集成学习 5](#_Toc514264501)

[2.3 常用分类器 7](#_Toc514264502)

[2.3.1 支持向量机 7](#_Toc514264503)

[2.3.2 神经网络 8](#_Toc514264504)

[2.4 本章小结 9](#_Toc514264505)

[第三章 ECOC算法介绍 11](#_Toc514264506)

[3.1 算法概述 11](#_Toc514264507)

[3.2 算法描述 11](#_Toc514264508)

[3.3 算法特点 14](#_Toc514264509)

[3.4 软编码 15](#_Toc514264510)

[3.5 本章小结 15](#_Toc514264511)

[第四章 软编码ECOC 17](#_Toc514264512)

[4.1 算法概述 17](#_Toc514264513)

[4.2 算法实现 18](#_Toc514264514)

[4.2.1 软值编码ECOC 18](#_Toc514264515)

[4.2.2 区间编码ECOC 18](#_Toc514264516)

[4.3 算法分析 21](#_Toc514264517)

[4.4 本章小结 23](#_Toc514264518)

[第五章 实验设计与结果分析 25](#_Toc514264519)

[5.1 实验设计 25](#_Toc514264520)

[5.2 实验结果及分析 27](#_Toc514264521)

[5.3 本章小结 38](#_Toc514264522)

[第六章 总结与展望 39](#_Toc514264523)

[6.1 总结 39](#_Toc514264524)

[6.2 工作展望 39](#_Toc514264525)

[参考文献 41](#_Toc514264526)

# Contents

[**Chapter 1 Preface 1**](#_Toc514264552)

[**1.1** **Introduction 1**](#_Toc514264553)

[**1.2** **The Structure of This Dissertation 3**](#_Toc514264554)

[**Chapter 2 Related Concepts 5**](#_Toc514264555)

[**2.1** **Multiclass Classification 5**](#_Toc514264556)

[**2.2** **Ensemble Learning 5**](#_Toc514264557)

[**2.3** **Commom Clssifiers 7**](#_Toc514264558)

[2.3.1 Support Vector Machine 7](#_Toc514264559)

[2.3.2 Neural Network 8](#_Toc514264560)

[**2.4 Summary 9**](#_Toc514264561)

[**Chapter 3 An Introduction on ECOC 11**](#_Toc514264562)

[**3.1 Summary on ECOC 11**](#_Toc514264563)

[**3.2** **Introduction of ECOC 11**](#_Toc514264564)

[**3.3 Features of ECOC 14**](#_Toc514264565)

[**3.4** **Soft Coding 15**](#_Toc514264566)

[**3.5 Summary 15**](#_Toc514264567)

[**Chapter 4 Soft Coding ECOC 17**](#_Toc514264568)

[**4.1** **Summary on the Algorithm 17**](#_Toc514264569)

[**4.2 Realization of Soft Coding ECOC 18**](#_Toc514264570)

[4.2.1 Soft Value Coding ECOC 18](#_Toc514264571)

[4.2.2 Soft Interval Coding ECOC 18](#_Toc514264572)

[**4.3** **Analysis on Soft Coding ECOC 21**](#_Toc514264573)

[**4.4 Analysis on Soft Coding ECOC 23**](#_Toc514264574)

[**Chapter 5 Experiments and Results 25**](#_Toc514264575)

[**5.1 Experiments Designing 25**](#_Toc514264576)

[**5.2 Experiments Results And Analysys 27**](#_Toc514264577)

[**5.3** **Summary 38**](#_Toc514264578)

[**Chapter 6 Conclusions and Feature Work 39**](#_Toc514264579)

[**6.1 Conclusions 39**](#_Toc514264580)

[**6.2 Feature Work 39**](#_Toc514264581)

[**References 41**](#_Toc514264582)

# 第一章 绪论

## 1.1 引言

分类问题一直是数据挖掘领域相关研究的重点问题之一，它的本质是以某种方式寻找出用于训练的数据的划分边界，从而可以分配未知类别数据在不同区域中，最后预测出这些数据的实际类别。解决分类问题的工具被称为分类器。研究分类问题的最终目标是实现拥有最佳分类性能的分类器。面对不同的分类问题，不同的分类器比如决策树（Decision Tree），K最近邻（k-Nearest Neighbor，KNN）等，根据它们的原理则会表现出不同的分类性能。分类问题主要可以分为二分类问题和多分类问题。在二分类问题中，分类器的分类性能在大多数情况下是相对比较优秀的，而在多分类问题中，因为出现了不同于二分类问题的高复杂性，只适用于二分类问题的分类器会失效，有些分类器如决策树，由于其原理依然可以适用于多分类问题，但对于如何提高分类性能，许多研究者为此展开了大量研究。

因为通常单个分类器不能正确地处理所有复杂地分类问题，所以实现该目标最有希望的方法是组合分类器，即集成学习。

集成学习通过将多个独立的分类器相结合，从而获得拥有更好分类性能的强分类器，而这些分类器应该具有可接受的准确度。在现有的研究中，已经有多种集成学习的方法可以产生优秀的强分类器，如boosting方法[1]，mixture of experts方法[2]和ECOC方法。每种方法都有其独到之处：boosting方法的重点是在每次训练中改变训练样本的分布，使得每个分类器可以专注于难以分离的点；mixture of experts则是尝试将样本空间划分成不同的子集，用这些子集去训练不同的分类器，使得这些分类器有不同的分工；而ECOC能把多分类问题分解为多个二分类问题，这些二分类问题分别对应一个二分类器，结合这些二分类器的结果做出决策[3]。本文讨论的重点是ECOC。纠错输出编码（Error Correcting Output Codes，ECOC）首先将问题分解，使得不同的二分类器只需要解决与它们对应的二分类问题，这些二分类器被称为基分类器，因为分类器在二分类问题的分类性能较为优秀，因此每个基分类器的输出结果的可靠性得到了增强。ECOC算法需要经过一个编码过程获得其编码矩阵，这个矩阵中的每一行称为码字，数据的每一类都对应着这个矩阵的每个码字。而每个基分类器都对应着编码矩阵的不同列。通过收集每个基分类器对待预测样本的输出值，可以计算出这个输出向量与每个码字的距离，这一距离可以理解成分配到与不同码字对应的类的置信度，藉此可以预测出样本所对应的类。

ECOC在真实数据集上取得了相对较好的成果[4-6]， 这也使得关于ECOC的研究一直被许多学者在进行着。ECOC的分类性能和编码矩阵是高度相关的，因此产生优质的编码矩阵非常重要，许多文献提出了各种方法。,E. L .Allwein等研究者在编码矩阵的构造上另辟蹊径，使得密集型编码矩阵（即只有+1,-1）变为稀疏型编码矩阵（包含+1，0，-1），扩展了ECOC在分解多分类问题的新思路[7]。O. Pujol等研究者根据二叉树的原理，提出了以互信息为依据，即考虑到样本数据特征的DECOC[8]。N. Hatami则尝试对编码矩阵进行优化，在原来的编码矩阵基础上，以每个基分类器对分类性能的贡献度为尺度，抛弃表现差的基分类器，以此提出了Thinned-ECOC[9]。最大化任意一对码字间的距离（对不同行）以及最小化基分类器之间的相关性（对不同列）。与此同时，ECOC的纠错能力也直接影响其分解性能。P. Simeone等研究者为增强纠错能力，把结果用拒绝域的形式展现，从而提出的ECOC多类别拒绝框架[10-12] 。

而在本文中讨论的偏标签分类问题，是一类特殊分类问题，无法直接使用上文中提到的任何算法，具体的表现在每个训练样本与一组候选标签相关联[14,15]。其中只有一个是对应于样本的真实标签，所有样本的真实标签隐藏在对应的候选标签集当中，普通的分类器都无法直接进行训练。针对这类特殊的分类问题，E. Hüllermeier 和 J. Beringer提出了相应的算法改进方案，对K最近邻和决策树算法进行改进，使得它们可以适用于偏标签数据分类问题[13]。无独有偶，M.L.Zhang等研究者提出了一种简单而有效的ECOC改进方法——PL-ECOC，同时也证明了它相比于集中成熟的偏标签分类算法具有更优越的性能[14]。

对于一般的分类问题，ECOC表现出纠错能力与泛化能力强，鲁棒性与可靠性高的特点，因此可以取得较好的结果。而目前对偏标签数据的分类需求越来越多的出现在现实应用中，无论是对被用户人为标注不同风格的绘画图像进行分类[16]还是关于自动化人脸识别的研究[17，18]，它们都需要应用偏标签数据的分类技术。在人脸估计年龄[19，20]，生物信息学[21]，对象分类问题[22]等研究中，偏标签分类技术都有成功的应用。因此，研究让本身分类性能优秀的ECOC可以更好地适用于偏标签数据分类问题，自然有其价值与意义存在。

本文的研究是建立在M.L.Zhang等研究者提出的PL-ECOC基础上，对这一简单而有效的算法进行改进。PL-ECOC使用的是与样本数据无关的完全随机的编码矩阵，而编码矩阵的构造与分类性能是高度相关的，在随机的条件下，分类性能不够稳定，无法完全地胜任偏标签数据地分类任务。而一些研究者根据数据复杂度来优化决策树，在UCI数据集上有不俗地表现[23]。受此启发而产生的以数据复杂度为依据的特征空间选择方法可以在一定程度上提升分类性能。但是，这些方法都无法直接作用于偏标签数据。受到Thinned-ECOC算法中编码矩阵构造的启发，依靠数据复杂度，那么可以进一步筛选出优秀的列用以组成最终更优质的编码矩阵。与此同时，虽然传统的特征选择方法无法直接作用，但是根据集成学习的思想，各个基分类器的分类性能会影响强分类器的性能，那么让每个基分类器的样本数据拥有较好的特征空间势必会对这个基分类器的分类性能产生影响。由于PL-ECOC算法与改进后的K最近邻算法最后结果均表现为置信度矩阵，那么以此作为先验信息，可以使得经过改进的PL-ECOC算法输出结果更加可靠。

## 1.2 论文组织结构

本文一共分成六章。本论文首先对偏标签数据、集成学习方法以及本文研究中所用到的支持向量机和适用于偏标签数据的K最近邻算法等相关概念进行介绍，并以此引出本文进行算法改进的对象——适用于偏标签数据的输出纠错编码（PL-ECOC），介绍其算法的特点与基本方法。接着，本文提出了可以改进PL-ECOC算法的三个方向，包括基于数据复杂度和样本完备性的编码矩阵构造、针对基分类器的特征空间选择方法和以K最近邻算法为先验信息优化最终输出结果的方式。介绍了它们的原理及实现后，分析了优缺点，并依靠实验来证明了这三种优化方式都会提升PL-ECOC的分类效果，且将这三种方式结合之后的算法在一些真实偏标签数据集上拥有更佳的分类性能，同时还分析了在这些数据集上分类性能得以提升的原因。最后，总结了本文的内容，并对后续的工作有所展望。

论文的具体安排如下：

第一章 简要介绍了多分类问题的实际应用和一些集成学习算法尤其是ECOC算法的概念，以及偏标签数据分类问题在实际数据挖掘应用中的重要性，展示了一些目前ECOC算法的相关研究成果和一些算法针对偏标签数据分类问题的改进，提出了改进方向，介绍了本文的研究背景、研究目标和意义。

第二章

第三章

第四章

第五章

第六章 总结了本论文的工作，对未来的改进方向提出展望

# 第二章 相关概念介绍

## 2.1 偏标签数据

偏标签数据与一般的数据不同，对于一般的数据，每一个数据只有一个真实标签会与之对应。形式上，定义为d维的实例空间，是一共有q类标签的标签空间，设是由m个样本构成的集合，对于X中的每个实例X，是由d维的特征向量的特征向量所表示的，而

## 2.2 集成学习

## 2.3 支持向量机

## 2.4 适用于偏标签数据的K最近邻算法

## 2.5 本章小结

# 第三章 PL-ECOC算法介绍

## 3.1 算法概述

## 3.2 算法描述

## 3.3 算法特点

## 3.4 软编码

## 3.5 本章小结

# 第四章 软编码ECOC

## 4.1 算法概述

## 4.2 算法实现

### 4.2.1 软值编码ECOC

### 4.2.2 区间编码ECOC

## 4.3 算法分析

## 4.4 本章小结

# 第五章 实验设计与结果分析

## 5.1 实验设计

## 5.2 实验结果及分析

## 5.3 本章小结

# 第六章 总结与展望

## 6.1 总结

## 6.2 工作展望

# 参考文献

1. Yoav Freund and Robert E. Schapire. 1996. Game theory, on-line prediction and boosting. In Proceedings of the ninth annual conference on Computational learning theory (COLT '96). ACM, New York, NY, USA, 325-332.
2. Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 1930-1939
3. T. G. Dietterich, G. Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes[J]. Journal of Artificial Intelligence Research, 1995, 2:263-286.
4. Hatami, N., Ebrahimpour, R., Ghaderi, R. (2008). ECOC-based training of neural networks for face recognition. In 3rd IEEE international conference on cybernetics and intelligent systems (CIS) (pp. 450–454).
5. Windeatt, T., & Ghaderi, R. (2001). Binary labeling and decision level fusion. Information Fusion, 2, 103–112.
6. Windeatt, T., & Ghaderi, R. (2003). Coding and decoding strategies for multi-class learning problems. Information Fusion, 4, 11–21.
7. Allwein, E. L., Shapire, R. E., & Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1, 113–141.
8. Escalera, S., Pujol, O., & Radeva, P. (2007). Boosted landmarks of contextual descriptors and forest-ECOC: A novel framework to detect and classify objects in cluttered scenes. Pattern Recognition Letters, 28, 1759–1768.
9. N. Hatami. Thinned-ECOC ensemble based on sequential code shrinking [J]. Expert Systems with Applications, 2012,39:936-947.
10. P. Simeone, C. Marrocco, F. Tortorella. A Framework for Multiclass Reject in ECOC Classification Systems[M]. Computer Science, 2007.
11. P. Simeone, C. Marrocco, F. Tortorella. Two Stage Reject Rule for ECOC Classification Systems [M]. Computer Science, 2011.
12. P. Simeone, C. Marrocco, F. Tortorella. Design of reject rules for ECOC classification systems[J]. Pattern Recognition, 2012, 45: 863-875.
13. E. Hullermeier and J. Beringer, “Learning from ambiguously labeled examples,” Intelligent Data Analysis, vol. 10, no. 5, pp. 419– 439, 2006.
14. Zhang M L , Yu F , Tang C Z . Disambiguation-Free Partial Label Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2017:1-1.
15. R. Jin and Z. Ghahramani, “Learning with multiple labels,” in Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 921–928.
16. L. Jie and F. Orabona, “Learning from candidate labeling sets,” in Advances in Neural Information Processing Systems 23, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Cambridge, MA: MIT Press, 2010, pp. 1504–1512.
17. Z. Zeng, S. Xiao, K. Jia, T.-H. Chan, S. Gao, D. Xu, and Y. Ma, “Learning by associating ambiguously labeled images,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 708–715.
18. T. Cour, B. Sapp, C. Jordan, and B. Taskar, “Learning from ambiguously labeled images,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 919–926.
19. F. Yu and M.-L. Zhang, “Maximum margin partial label learning,” Machine Learning, vol. 106, no. 4, pp. 573–593, 2017.
20. M.-L. Zhang, B.-B. Zhou, and X.-Y. Liu, “Partial label learning via feature-aware disambiguation,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 1335–1344.
21. F. Briggs, X. Z. Fern, and R. Raich, “Rank-loss support instance machines for MIML instance annotation,” in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012, pp. 534–542.
22. L.LiuandT.Dietterich,“Aconditional multinomialmixturemodel for superset label learning,” in Advances in Neural Information Processing Systems 25, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 2012, pp. 557–565.
23. Tan J, Chua K S, Zhang L. Algorithmic and Complexity Issues of Three Clustering Methods in Microarray Data Analysis[M]// Computing and Combinatorics. Springer Berlin Heidelberg, 2005:74-83.