密级： 保密期限：

北京郵電大學

硕士研究生学位论文

题目：SVM 分类器置信度的研究

学　　号：076380

姓　　名：赵行

专　　业：模式识别与智能系统

导　　师：何华灿 教授

学　　院：计算机学院

2010 年 3 月 10 日

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：2010.3.18

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密在__年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名：_____ 日期：2010.3.18

导师签名：_____ 日期：2010.3.19

SVM 分类器置信度的研究

摘 要

当今，由于海量数据的形成，迫切需要将这些数据转换成有用的信息和知识，促进了数据挖掘的应用，使这一技术迅速得到发展和完善。数据挖掘是数据库、人工智能和统计学等学科的研究热点领域之一。而分类技术作为数据挖掘技术的一个重要研究方面，受到研究者们的广泛关注。

支持向量机（SVM）是最重要的分类技术之一，也是近年来机器学习领域的研究热点，是借助最优化理论解决分类问题的有力工具，在许多实际分类应用中都表现了良好的性能。

本文对 SVM 分类器分类结果的置信度评估及决策修正算法进行了研究，利用直接得到的观察量来反映识别结果之间的相对可靠性，共设计了四种算法。通过实验证明了其中一种算法为最佳算法。在该算法中，我们在分类器预测阶段获取待测样本和最优分类超平面的距离，并且计算待测样本的 j 个近邻训练样本与待测样本经 Libsvm 判断的初始分类结果同属一类的概率。对于给定的拒识率，该算法拒识并修正置信度小于相应置信度阈值的样本分类结果。实验结果证明此置信度评估及决策修正算法能够很好地提高分类中常用的 Libsvm 分类器的性能，并且该算法具有相当的稳定性。

关键词：支持向量机 置信度 阈值 拒识

RESEARCHES ON ALGORITHM FOR CONFIDENCE EVALUATION OF SVM

ABSTRACT

At present, the emergence of vast amounts of data, urgently need to be converted into useful information and knowledge. This promoted application of data mining, and this technology rapidly developed and improved. Data mining is a active research field involving disciplines such as databases, artificial intelligence and statistic. Among them, as an important aspect of data mining technology, classification technology has always been concerned about by researchers.

In recent years, support vector machine is one hot technology of machine learning research fields, and also the most important classification technologies, a powerful tool to solve classification problems using optimization theory. It showed good performance in many practical applications.

This paper focuses on the research of algorithm to estimate confidence measure and decision modification of support vector machine

(SVM). Totally 4 algorithms using parameters to reflect the confidences of recognition results are presented in this paper, and experiment shows that one algorithm of them is the best algorithm. This algorithm computes the distance from testing sample to the optimal hyperplane of SVM, and the probability that the testing sample and its k nearest neighbors belong to the same class as the decision of Libsvm for the testing sample. The algorithm rejects the classification results of samples whose confidence measures are smaller than the threshold corresponding to a given rejection rate. Experiments show that the performance of the Libsvm classifier has been well improved using this algorithm.

KEY WORDS: Support vector machine, Confidence measure, Threshold, Rejection

目录

第一章 绪论

1.1 选题研究背景

1.1.1 引言

在社会迅速发展的今天，随着计算机和因特网的不断发展和广泛普及，人们接触的事物越来越多，人们获得的数据也正以前所未有的速度急剧增加。

海量数据的形成，人们迫切需要从海量数据中快速获取有效信息。但是，如此庞大的数据中蕴含的信息和知识亟待人们去获取发现。人们面对的问题是被大量数据淹没了，而人们分析处理它们的能力以及从中获取知识的能力都与之存在着相当大的差距[1]。面对如此的挑战，如何开发有效的挖掘方法，已成为众多研究者们关注的焦点。因此，一个新的研究领域——数据挖掘技术应运而生，并且越来越显示出其强大的生命力，这一技术迅速得到广泛的发展和完善。目前它已成为人工智能、统计学、数据库等领域研究的热点之一[2]。而分类技术作为数据挖掘技术的一个重要研究方面，是一种重要的数据分析技术，受到研究者们的广泛关注。

分类算法研究的过程中产生了很多好的解决方法，其中本论文中研究的支持向量机方法(Support Vector Machine，简称 SVM)就是其中一个有效的分类方法。本文通过对支持向量分类结果的置信度进行研究，并最终提高了支持向量机分类结果的准确率。

本章简要概述了数据挖掘技术，SVM 分类器的应用现状，研究分类器置信度的意义，并且说明了本论文的主要工作。

1.1.2 数据挖掘概述

数据挖掘(Data Mining，简称 DM)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取出潜在的、可信的、新颖的、有价值的信息和知识(模型或规则)的过程，是一类深层次的数据分析方法[3]。数据挖掘综合利用了来自模式识别、机器学习、统计学、数据库以及人工智能等各学科的知识技术，并且已在经济、商业、金融、天文等行业得到了广泛而成功的应用。

许多人把数据挖掘视为另一个常用术语，也就是数据库中的知识发现(Knowledge Discovery in Database，简称 KDD)的同义词。文献 4 认为知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。知识发现将信息变为知识，从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展作出贡献。而另一些人把数据挖掘视为数据库中知识发现过程中的一个环节。总之，数据挖掘的定义与术语数据库中的知识发现密切相关。

数据挖掘的任务主要是相关性分组或关联规则(Affinity grouping or association rules)、聚类分析(Clustering)、分类(Classification)、描述和可视化(Description and Visualization)等[5]。

其中，分类是数据挖掘的一项非常重要的任务，分类技术是数据挖掘技术的一个重要研究方面，其工作方法是通过分析已知分类信息的历史数据总结出一个预测模型——分类器(也称作分类函数，分类模型，分类规则，假设)。该分类器能够依据数据属性将数据映射到不同类别中。这样我们就可以利用该函数来分析已有数据，并预测新数据的趋势，即预测新数据属于哪一个类。现实中的诸多问题都可以转化为分类问题，因而数据挖掘分类技术的潜在应用十分广泛。例如，银行信用分类就是典型的分类挖掘问题[6]；也可以通过建立分类模型，运营商的客户群进行客户分类，进而对不同用户推荐其可能感兴趣的增值业务等。下节将详细介绍 SVM 分类器的应用现状。

1.2 SVM 分类器的研究现状

V. Vapnik等人从六、七十年代开始致力于统计学习理论这方面研究，到九十年代中期，随着其理论的不断发展和成熟，统计学习理论开始受到越来越广泛的重视[7][8][9]。

统计学习理论是专门研究小样本情况下机器学习规律的理论，建立在一套较坚实的理论基础之上，为解决有限样本学习问题提供了一个统一的框架[10]。在这一理论基础上发展了一种新的通用学习方法——支持向量机，支持向量机是由Cortes & Vapnik在1995年首先提出来的。支持向量机是最重要的分类技术之一，也是近年来机器学习领域的研究热点，是借助最优化理论解决分类问题的有力工具。SVM建立在一套坚实的理论基础之上，在解决有限样本学习问题时表现出优异的性能。SVM在所有的分类面内，寻找的是最优超平面，这个最优超平面能够使得两类的分类间隔最大。

由于SVM算法的潜在应用价值，国内外许多研究者都致力于此方面的研究，

近几年出现了许多发展和改进的SVM算法[11][12][13]。

SVM 方法在理论上具有突出的优势，SVM 对经验的依赖较小，能够获得全局最优解以及良好的泛化性能，随后的几年内，有关 SVM 的应用研究得到了很多领域的学者的重视，众多的研究者转向了对 SVM 的研究。在人脸检测、验证和识别、说话人/语音识别、文字/手写体识别、图像处理及其它应用研究等方面取得了大量的研究成果，对推动支持向量机的研究具有重要意义。

（1）人脸图像识别

SVM 应用于人脸检测是由 Osuna 最早进行的研究，并取得了较好的实验效果。Osuna 将非线性 SVM 分类器用以完成人脸与非人脸的分类[14]。

（2）手写字识别

SVM 方法在理论上具有突出的优势，贝尔实验室率先在美国邮政手写数字库识别研究方面应用 SVM 方法取得了较大的成功，表明了 SVM 的优越性能[15]。

（3）语音识别

文献 16 和文献 17 很好地将 HMM 和 SVM 进行组合，解决了说法人识别的问题。说话人识别属于连续输入信号的分类问题，由于 HMM 适于处理连续信号，SVM 适于处理分类问题的特性；根据两者不同的侧重点，使其组合获得了很好的效果。

（4）指纹识别

指纹识别是近年来研究的热点问题，它是网络安全的一项重要研究课题，文献 18 尝试用支持向量机来解决指纹的识别问题，取得很好的效果。

（5）文本分类

近几年将 SVM 算法用于文本分类中的应用非常广泛，文献 19，20 都针对文本分类的问题研究了主动学习方法，有效地减少学习所需的样本数量，很好地解决了小规模标注样本集的分类问题。

（6）垃圾邮件过滤

马莉[21]设计了一个嵌入到 OUTLOOK2000 中使用 SVM 作为分类器的垃圾邮件过滤系统。通过使用支持向量机 SVM 作为分类算法，为垃圾邮件或者正常邮件建立一个分类器，对邮件进行过滤。

（7）医学应用

王朝勇[22]把 SVM 应用到药品成分分析和肿瘤性质诊断这两个领域中。关于肿瘤性质诊断的实验结果证实了该文种提出的混合特征处理方法的有效性，关于药品成分分析的实验结果验证了 SVM 在相关领域应用的可行性与有效性。

（8）图像处理

张磊[23]研究了图像的检索问题，SVM 算法在有限训练样本下有较强的泛化

能力，检索结果无论从查全率和查准率两方面较传统方法都有较大的提高。肖靓通过实验分析了特征选取对向量机性能的影响[24]，发现综合特征有利于分类效果的提高，建立了一个基于SVM的图像分类实验平台，验证了前述理论研究的结果。付岩等人[25]以自然图像领域为例，使用支持向量机(SVM)学习自然图像的类别，学习到的模型用于自然图像分类和检索.实验结果表明作者的方法是可行的。

（9）其它研究

由于SVM的优越性，其应用研究的开展已经相当广泛，在越来越多的方面取得优异的应用成果，其应用领域日趋广泛，例如前文提到的银行个人客户信用分类、运营商客户群分类等，SVM还涉及了智能交通等诸多领域。

1.3 研究分类器置信度的意义

所谓置信度，也叫置信水平。它是指特定个体对待特定命题真实性相信的程度。也就是概率是对个人信念合理性的量度。概率的置信度解释表明，事件本身并没有什么概率，事件之所以指派有概率只是指派概率的人头脑中所具有的信念证据。置信度也称为可靠度，或置信水平、置信系数[26]。

林晓帆等人[27]提出，对于任何一个分类器，我们不仅总体上希望该分类器的分类准确率尽可能地高，还希望能较为准确地估计每一个测试样本分类结果的准确性，也就是每个识别结果的置信度。因此分类器的置信度也应当是一个需要研究的重要参量，因为它在决定拒识门限和多分类器集成中起着关键作用。分类器置信度的研究已经得到了广泛的应用，拒识区域的选择、识别率的估计和多分类器的集成[28]、手写数字识别和脱机手写汉字识别的实际应用都验证了所提出的理论和方法[29]。

置信度的主要用途有[27]：（1）能够为拒识提供依据。在一些应用中，我们不仅希望识别率尽可能高，同时也希望误识率尽可能低，这就得通过拒识部分样本实现。因此，需要决定拒识的样本为哪一部分，如果我们选择拒识的总是识别置信度最低的样本，那么在拒识率一定的情况下，系统的误识率就会获得最大幅度的下降。（2）为多方案集成提供根据。现在，人们已不再满足于简单的分类器决策，而是希望能充分利用每个分类器的各种信息。其中，分类器的置信度就是一种重要信息。

目前现有的分类器识别结果的置信度评估算法，绝大多数都是针对最近邻、神经网络、贝叶斯等分类器进行的分析，对SVM分类器置信度的研究则比较少。

1.4 本论文主要工作

本论文设计算法对 SVM 分类器分类结果的置信度进行评估，对于该 SVM 分类器我们使用的是 Libsvm 工具包，根据评估出的 SVM 分类结果的置信度，对低置信度也就是高风险的分类结果进行修正，进而达到提高 Libsvm 分类结果的目的。

文章的其余部分按下面的结构组织：

第二章了介绍了朴素贝叶斯分类器、最近邻分类器、KNN 分类器、神经网络分类器的基本原理及其相应的置信度评估算法，得出可以采用直接得到的观察量反映识别结果之间的相对可靠性的结论。

第三章首先说明了 SVM 分类器的基本原理，并介绍了一种 SVM 分类器的置信度评估算法，接着详细说明了本文所提出的四种 SVM 分类器的置信度评估及决策修正算法的设计与实现。

第四章是本论文的实验设计和结果分析。首先简要说明了实验数据的准备工作，接着是多组实验及结果分析，得出最终的 SVM 分类器置信度评估及分类结果修正算法，最后进行总结。

第五章对本课题的研究进行总结与展望。

第二章 现有的分类器的置信度算法研究

本章将详细介绍朴素贝叶斯分类器、最近邻分类器、K近邻分类器和多层前向神经网络的基本原理及其分类结果的置信度评估算法。

2.1 朴素贝叶斯分类器

2.1.1 朴素贝叶斯分类原理

先验概率是指根据历史上的资料或者主观判断所确定的各事件发生的概率，反应了各种“原因”发生的可能性，一般是以往经验的总结且在实验之前已经知道[30]。该类概率没能经过实验证实，属于检验前的概率，所以称之为先验概率。先验概率通常分为两类，利用过去的历史资料计算得到的概率称之为客观先验概率；是指当历史资料无从取得或者历史资料不完全时，只能凭借人们的主观经验来判断取得的概率称之为主观先验概率[31]。

后验概率一般是指利用贝叶斯公式，结合调查等方式获取了新的附加信息，对先验概率进行修正后得到的更符合实际的概率[32]。它反应了试验之后对各种“原因”发生的可能性大小的新知识。

贝叶斯公式也叫后验概率公式，有着广泛的用途。设先验概率为 $P(B_i)$，调查所获得的新附加信息为 $P(A \mid B_i)$，其中 $i = 1, 2, \cdots, n$，则贝叶斯公式计算的后验概率为[32]

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{k=1}^{n} P(B_k)P(A \mid B_k)} \qquad \text{式（2-1）}$$

贝叶斯网络能发现数据间的潜在联系。在这个网络中，用结点表示变量，用有向边来表示变量之间的依赖关系。如果在贝叶斯网络中把其中代表类别变量的结点作为根结点，其余所有变量都作为它的子结点时，贝叶斯网络就变成了分类器。

2.1.2 朴素贝叶斯分类器的置信度评估算法

王利民提出[32]，在数据挖掘中，通常感兴趣的是在给定训练数据 D 时，确定假设空间 H 中的最佳假设。贝叶斯法则基于假设的先验概率、给定假设下观察到不同数据的概率、以及观察的数据本身，提供了一种计算假设概率的方法。

以文本分类应用中的朴素贝叶斯分类为例，文献 33 介绍了贝叶斯分类器通过类别的先验概率 $P(C_j)$和词的分布来计算未知文本属于某一类别的概率：

$$P(Cj|D)=P(Cj)*P(D|Cj)/P(D) \qquad \text{式（2-2）}$$

其中，$P(C_j|D)$为文本 D 属于类 C_j 的概率，$P(D|C_j)$为类 C_j 中含有文本 D 的概率。在所有 $P(C_j|D)$ $(j=1, 2, \cdots, m)$中，若 $P(C_k|D)$值最大，则文本 D 归为 C_k 类。由于 $P(D)$是常数，因此将要求解 $P(C_j|D)$的问题转换为只要求解 $P(C_j)P(D|C_j)/P(D)$。假设文本中词的分布是条件独立的，则

$$P(Cj|D)=P(Cj)*P(D|Cj)/P(D)=\frac{P(C_j)}{P(D)}\prod_{j=1}^{M}P(d_j|C_j) \quad \text{式（2-3）}$$

其中

$$P(C_j)=C_j\text{中文本个数}\Big/\text{总文本个数}$$

$$P(d_j|C_j)=d_j\text{在类}C_j\text{中出现的次数}\Big/C_j\text{中所有词的个数} \qquad \text{式（2-4）}$$

根据文献 32，在数据挖掘，我们感兴趣的就是 $P(C_j|D)$，因为它反映了在看到训练数据 D 后 C_j 成立的置信度。

2.2 最近邻分类器

2.2.1 最近邻分类器理论基础

最近邻分类器将与测试样本最近邻样本的类别作为决策的结果。

郭亚琴如下介绍了最近邻分类器[34]：

假设有 c 个模式类 $\omega_1, \omega_2, \ldots, \omega_c$，对于待识别样本 x，计算它到模式类 ω_i 的所有样本 $x_{is}(i=1, 2, \ldots, c; s=1, 2, \ldots, n_i)$的距离，即：

$$d_i(x)=\min_s\|x-x_{is}\| \qquad s=1, 2, \ldots, n_i \qquad \text{式（2-5）}$$

常用的最近邻分类器就是根据式（2-4）定义的距离 $d_i(x)$，对 x 进行分类决策，决策规则如下：

$$e(x)=arg\left[\min_{1\le i\le c} d_i(x)\right] \qquad \text{式（2-6）}$$

$e(x)$就判断为 x 所属的类别。

2.2.2 最近邻分类器置信度评估

对于基于距离的模式识别分类器，如最近邻分类器等这类分类器，测试样本距各类训练样本的距离与置信度有着密切的联系[35]。

结合对最近邻分类器置信度的理论分析和应用实践[36][37]，证明了以下两式均是对最近邻分类器的置信度：

$$f_1(x)=1-d_1(x)/d_2(x) \qquad \text{式（2-7）}$$

$$f_2(x)=(d_2(x)-d_1(x))/(d_1(x)+d_2(x)) \qquad \text{式（2-8）}$$

式中：$d_1(x)$是测试样本 x 与最近的训练样本 x_m 间的距离，$d_1(x)$是测试样本 x 与训练样本中与 x_m 属于不同类别的其它训练样本的最近距离。由于 $d_1(x)$ 不可能比 $d_2(x)$大，所以置信度置于 0 和 1 之间。

2.3 K 近邻分类器

2.3.1 k 近邻分类器理论基础

k 近邻法是由 Cover 和 Hart 于 1968 年提出的[38]，k 近邻分类器是一种懒散的学习方法，它的基本思想是：对于给定的测试样本 x，首先找到训练集中与之距离最近(最相似)的 k 个样本，然后根据这 k 个样本所属的类别来判定待测样本 x 的类别。简单地说，系统对测试样本在训练集中找到最近的 k 个近邻，看这 k 个近邻中多数属于哪一类，就把待识别的样本归为那一类[3]。

k 近邻分类器相似值的大小可以使用欧拉距离，或是余弦相似度等多种形式衡量。最相似的 k 个样本按其和待分类样本的相似度高低对类别值进行加权平均，从而预测待测样本的类别[39]。

当我们利用 k 近邻分类器进行样本分类时，对于二分类问题，通常采用简单多数原则确定测试样本 x 的类别，即以 k 近邻中出现次数最多的类别作为样本 x 的分类类别。对于多分类问题，通常用下式计算测试样本 x 与各个 C_j 类别的相

似性[40]：

$$y(x, c_j) = \sum_{d_j \in knn(x)} sim(s, d_j)\, y(d_j, c_j) - b_j \qquad \text{式（2-9）}$$

其中，公式 $sim(s, d_j)$用来计算相似度，$y(d_j, c_j)$为类别属性函数。如果 d_j 属于类 c_j，则函数值为 1，否则为 0。b_j 为阈值，可以通过训练取得。如果 $y(x, c_j) > 0$，k 近邻分类器判别测试样本 x 属于类别 c_j，否则判决其不属于类别 c_j。

2.3.2 KNN 分类器置信度评估算法

薛磊等人[41]通过对 KNN 算法的分析，将 KNN 分类器的置信度定义为：

$$T = d_m(x)/k \qquad \text{式（2-10）}$$

式中，k 为测试样本的近邻点的个数，$d_m(x)$为测试样本通过 KNN 判断所属类别包含的近邻训练样本点的个数，k 与 $d_m(x)$相差越小，置信度 T 越大，该测试样本的分类结果可靠性就越高。

2.4 神经网络

2.4.1 神经网络理论基础

人们对神经网络（Neural network，NNet）的研究源于物理学、心理学和神经生理学的跨学科研究[42]。神经网络是一组连接的输入输出单元，并且每个连接由一定的权值相连。在学习阶段，根据目标输出，通过不断调整神经网络的相连权值，得到满足误差要求的网络实际输出，记录此时的权值参数，进而预测测试样本的类别。神经网络通常由输入层、输出层和隐藏层组成，输入层的神经元个数等于训练样本特征项的个数，输出层即为分类判决层，其神经元个数等于样本的类别数[43]。

在神经网络模型中，多层前向神经网络(Multi - Layer Feedforward Neural Networks - MLFNN)是应用得最为广泛的一种网络[44]。它一般由输入层、输出层和至少一个隐藏层组成，各层均有一个或多个神经元，相邻两层间神经元都由可调权值相连，但同层神经元之间不互相连接，并且整个网络不存在反馈。信息由输入层依次经隐层向输出层传递[45]。

BP 网络是 D . E . Rumelhart 和 J . L . McClelland 等人 1986 年提出的一种多层前馈性型的误差反向传播（ Back Propagation ）算法网络[46]。BP 算法根据学习

的误差，把学习的结果反馈到隐含层的神经元，修正调整各层之间的连接权值，从而达到预期的学习目的，使得网络预测与实际类之间的均方误差最小，解决了多层网络的学习问题[43]。

BP 算法基本步骤[43]：

Step 1 ：初始化神经网络各层的权值及神经元节点的阈值。

Step 2 ：向前传播输入。对每一样本，计算隐藏层和输出层每个单元的净输入和输出。

Step 3 ：后向传播误差。通过更新权值和偏置以反映网络的预测误差。

Step 4 ：终止条件：1、更新权值较小；2、正确分类的样本百分比；3、超过预先指定的训练周期。

2.4.2 多层前向神经网络分类器的置信度评估算法

在多层前向神经网络分类器的置信度估计中，当使用均方误差或库尔贝克(Kullback) 鉴别熵做代价函数时，多层前向神经网络输出的期望值是各个类别的后验概率[35][47]。

设 o_i 是与 ω_i 类 对应的神经网络的输出，则：

$$E\{o_i\} = P(\omega_i | x) \qquad \text{式（ 2-11 ）}$$

判决时取与最大输出对应的类别，因此

$$c(x) = E\{max\, o_i\} \qquad \text{式（ 2 - 12 ）}$$

也就是可以用神经网络的最大输出做置信度。

2.5 小结

本章详细介绍了朴素贝叶斯分类器、最近邻分类器、K 近邻分类器和多层前向神经网络的基本原理及其分类结果的置信度评估算法。通过对以上几种分类器的分析，我们发现，正如林晓帆等人[35]提出，例如最近邻分类器中的距离、K 近邻分类器中测试样本通过 KNN 判断所属类别包含的近邻训练样本点的个数等这些能直接得到的观察量，往往能反映识别结果之间的相对可靠性。

文献 35 阐述了广义置信度的思想。x 为从输入的待测样本提取的特征向量，x 的真实类别为 $\omega(x)$，模式分类器 S 对 x 的判决为类别 $e_s(x)$，则将判决 $e_s(x)$ 正确的概率定义如下：

$$p(e_s(x)=\omega(x))=c_s(x) \qquad \text{式（2-12）}$$

$c_s(x)$即为S在特征向量空间内点x处的置信度，识别置信度反映的是分类器S在特征向量空间某点的判决可信度。同时，若存在函数$f_s(x)$和一个单调递增函数$g(.)$，使得：

$$f_s(x)=g(c_s(x)) \qquad \text{式（2-13）}$$

则称$f_s(x)$为S的广义置信度。从上式可以看出，置信度$c_s(x)$是广义置信度的一个特例，一个分类器的置信度是唯一的，但广义置信度不是唯一的。因为广义置信度可以通过将任何一个单调递增函数作用在置信度来得到。可以说，广义置信度是[0，1]范围内的一种值域，如果$f_s(x_1)>f_s(x_2)$，则表示该分类器S在x_1处比在x_2处更可靠些。尽管不知道分类器在这两点实际判决正确的概率，但是广义置信度可以作为一种相对度量，因为分类器在这两点上判决的置信度的高低大小是唯一的。

因此，在本文的SVM分类器的置信度评估算法中，我们也利用直接得到的观察量来反映识别结果之间的相对可靠性。在本文的算法设计中，我们对待测样本经SVM判决的结果的广义置信度进行评估，拒识置信度相对较低的分类结果，接受置信度高的分类结果。

第三章 SVM 分类器置信度评估算法设计

3.1 支持向量机理论

支持向量机是近年来机器学习领域的研究热点，以统计学习理论为基础，避免了传统分类算法中样本无限大的问题，在解决有限样本学习问题时表现出优异的性能，具有很好的泛化性能，目前已成功应用于许多领域。

3.1.1 引言

SVM 方法建立在统计学习理论的 VC 维理论和结构风险最小化(Structural Risk Minimization，简称 SRM)原理基础之上，避免了局部最优解，并巧妙地克服了“维数灾难”，在解决小样本、非线性和高维输入空间等分类问题中表现出许多特有的优势[48]。因其独特的优势和出色的学习性能，显示出广泛的应用前景和重要的研究价值，已经成为一种备受关注的分类技术。这种技术具有坚实的统计学理论基础，并在许多实际应用（如手写数字的识别、文本分类等）中展示了优异的实践效用[49]。下面将做详细介绍。

3.1.2 统计学习理论

传统的分类算法的研究都是在样本无限大的理论前提下进行的，因此提出的各种方法只有在样本数趋于无穷大时其性能才有理论上的保证。然而，在实际应用中，样本数目通常是有限的，所以很多传统方法都难以取得理想的效果[50]。Vapnik 等人在 20 世纪 60 年代开始研究有限样本情况下的机器学习问题，20 世纪 90 年代，有限样本情况下的机器学习理论研究逐渐成熟起来，形成了一个较完善的理论体系——统计学习理论 (Statistical Learning Theory，简称 SLT)，该理论开始受到越来越广泛的重视。统计学习理论是建立在一套较坚实的理论基础之上的，是一种专门的小样本统计理论，它是在经验风险最小化 (Empirical Risk Minimization，简称 ERM)等研究的基础上发展起来的，其主要内容包括 VC 维和结构风险最小化(Structural Risk Minimization，简称 SRM)原理及相关定理，

系统地研究了有限样本情况下的统计模式识别和机器学习问题的原理与方法的理论[51]。

在统计学习理论中，机器学习问题可以概括为：假设输入 x 与输出变量 y 之间存在一定的未知的依赖关系，通过联合概率 $P(x,y)=P(x)P(y|x)$ 表示，机器学习的问题就是根据 n 个独立同分布(independently drawn and identically distribute)的观测样本[48][52]

$$(x_1,y_1),(x_2,y_2),...,(x_n,y_n),x_i\in X\subset R^n,y_i\in Y=\{1,...,k\} \qquad 式(3-1)$$

学习到一个假设 $H=f(x,\omega)$ 作为预测函数，其中 $\omega\in O$ 是预测函数的广义参数，它对 $P(x,y)$ 的期望风险（既统计学习的实际风险）$R(\omega)$ 是：

$$R(\omega)=\int L(y,f(x,\omega))dP(x,y) \qquad 式(3-2)$$

其中 $L(y,f(x,\omega))$ 表示用 $H=f(x,\omega)$ 预测结果对真实值 y 进行预测所造成的损失。不同类型的学习问题，有不同形式的损失函数。

1. 经验风险

在实际的机器学习问题中，由于联合概率 $P(x,y)$ 是未知的，因此也就难以直接计算实际风险 $R(\omega)$，只能利用已知训练样本的信息，用算术平均代替实际风险 $R(\omega)$，于是使用算术平均

$$R_{cmp}(\omega)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i,\omega)) \qquad 式(3-3)$$

来逼近式(3-2)定义的期望风险 $R(\omega)$。由于 $R_{cmp}(\omega)$ 是用已知的训练样本（即经验数据）定义的，因此称作经验风险。

2. 经验风险最小化原理

通过求经验风险 $R_{cmp}(\omega)$ 的最小值代替求期望风险 $R(\omega)$ 的最小值，就是所谓的经验风险最小化原则，简称 ERM 原则。

许多经典的分类算法如神经网络、决策树等，实际上都是在经验风险最小化原则下提出的。但使用经验风险最小代替期望风险最小并没有可靠的理论依据、相反却存在缺陷[53]：

1）$R_{cmp}(\omega)$ 和 $R(\omega)$ 都是 ω 的函数，概率论的大数定理说明，只有当样本数趋于无穷多时，才有 $R_{cmp}(\omega)$ 在概率意义上趋近 $R(\omega)$；

2）$R_{cmp}(\omega)$ 最小时的参数值 ω^*_{cmp} 与 $R(\omega)$ 最小时的参数值 ω^* 无法保证是同一值，更不能保证 $R_{cmp}(\omega^*)$ 能够趋近于 $R(\omega^*)$。

因此，我们可以看到：经验风险 $R_{cmp}(\omega)$ 的最小，并不意味着实际风险 $R(\omega)$ 的最小。某些情况下，当训练误差过小反而会导致推广能力的下降，导致实际风

险的上升，这就是某些机器学习方法中出现的所谓过学习问题。

3. VC 维[50]

统计学习理论定义了一系列有关函数集学习性能的指标，用以研究学习过程一致收敛的速度和推广性，VC 维就是其中重要的一个指标，其概念的直观定义是：对于一个指示函数集，如果存在 h 个样本能够被函数集里的函数按照所有可能的 2^h 种形式分开，则称函数集能够将 h 个样本打散。函数集的 VC 维就是它能够打散的最大样本数目 h。如果函数能够打散任意数目的样本，则称函数集的 VC 维是无限的。

VC 维描述了组成学习模型的函数集合的容量，刻画了此函数集合的学习能力。VC 维越大，函数集合越大，其相应的学习能力就越强，但机器容量也随之增大，会导致学习过程更加复杂。

4. 结构风险最小化原理

统计学习理论系统地研究了有限样本条件下有关经验风险与期望风险之间关系等问题，并据此提出了结构风险最小归纳原理，克服了以往经验风险最小化原则的缺点。

对于经验风险 $R_{cmp}(\omega)$和实际风险 $R(\omega)$，统计学习理论用这样的定理评价它们之间的关系：对于二分类问题中的指示函数集 $H=f(x,\omega)$中的所有函数，经验风险 $R_{cmp}(\omega)$和实际风险 $R(\omega)$之间至少以至少 $1-\eta\ (0\le\eta\le 1)$的概率满足这样的关系[52]：

$$R(\omega)\le R_{cmp}(\omega)+\varphi(h/l) \qquad \text{式(3-4)}$$

其中 h 是函数 $H=f(x,\omega)$的 VC 维，n 是样本数，$f(h/n)=\sqrt{\frac{h(\ln(2n/h+1)\ln(\eta/4))}{n}}$ 被称为置信风险，。

这一结论从理论上说明了学习机器的实际风险由两部分组成：一部分是经验风险 $R_{cmp}(\omega)$，另一部分为置信范围 $f(h/n)$，它和机器的 VC 维 h 及训练样本数 n 有关。

根据公式(3-4)，如果训练样本数目 n 不变，则控制风险 $R(\omega)$的参量有两个：ω 与 h。一般的学习方法是基于 $R_{cmp}(\omega)$最小，满足对已有的训练数据的最佳拟合，在理论上可以通过增加算法的规模使得 $R_{cmp}(\omega)$不断降低以至为 0。但这样使得算法的复杂度增加，VC 维 h 增加，从而 $f(h/l)$增大，期望风险 $R(\omega)$和经验风险 $R_{cmp}(\omega)$之间的差别越大，导致实际风险 $R(\omega)$增加，对测试集的预测能力下降，这是此类学习算法产生过学习现象的原因[51]。

根据式(3-4)的理论依据，统计学习理论提出了一种新的策略：首先把函数

集 $S=\{f(x,\omega),\omega\in\Omega\}$ 构造为一个函数子集序列，并将相应各个子集按照VC维的大小排列，这样在同一个子集中的置信范围就相同[48]。在每个子集中寻找最小经验风险，在子集间折衷考虑经验风险和置信范围，选择最小经验风险与置信范围之和最小的子集，也就能够在获得的学习模型经验风险最小的同时，置信风险尽可能小，学习模型的推广能力尽可能大，从而取得最小化的期望风险，这种思想称作结构风险最小化，即SRM准则。

3.1.3 支持向量机[48][49]

支持向量机是建立在VC维和结构风险最小化基础上的一种新的机器学习方法，它是在有限样本的基础上，在训练复杂度和学习能力之间寻求折中，以期望获得较好的推广能力，这种方法被广泛用于模式识别领域[54][55]。

1）线性可分的标准最优分类面

SVM是从线性可分情况下得最优分类面提出得，其基本思想可用图3-1所示的用某特征空间的超平面对给定训练数据集作二分类说明。

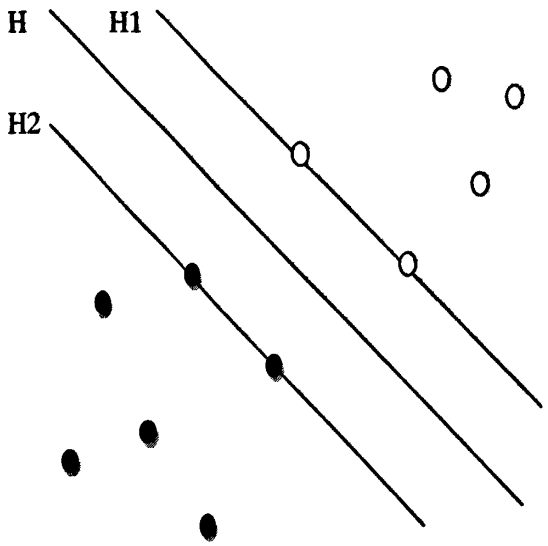

图3-1 支持向量机的示意图

如图3-1所示，实心点和空心点代表两类样本，H为分类超平面，H1，H2分别为过各类中离超平面最近的样本且平行于超平面的平面，它们之间的距离叫做分类间隔（Margin）。最优分类超平面就是要求不但能将两类正确分开，即训练错误率为0，而且还要使分类间隔最大。

在线性可分的情况下，在特征空间 R^n 中可以构造多个分割平面（如：H1，H2，……），分类超平面的方程被定义为：

$$w^T x+b=0 \qquad \text{式}(3-5)$$

同时，这个分类面能将两类（ P，N ）无误差的完全分开，既满足

$$\begin{aligned} w^T x_i + b \geq 1, \quad & \text{for all } x_i \in P \\ w^T x_i + b \leq -1, \quad & \text{for all } x_i \in N \end{aligned} \qquad \text{式(3-6)}$$

此时的分类间隔

$$dist = \frac{2}{\|\omega\|} \qquad \text{式(3-7)}$$

满足约束条件，并且使分类间隔 *dist* 最大的超平面就叫做最优分类超平面，H1，H2 上的训练样本点就称作支持向量。

根据以上分析，求解最优超平面就相当于，在式(3-6)的约束条件下，求间隔最大即式 (3-7)的最大值，也就是：

$$\begin{aligned} &\underset{\omega,b}{\text{Minimize}} \ \ \Phi(\omega) = \frac{1}{2}\|\omega\|^2 \\ &\text{subject to } y_i(\omega^T x_i + b) \geq 1, \quad i = 1,2,.....,l \end{aligned} \qquad \text{式(3-8)}$$

它是一个线性约束下的凸二次式优化问题，可根据 Lagrange 方法求解，最终得到的判别函数是

$$f(x) = \text{sgn}(\sum_{i=1}^{l} y_i \lambda_i^* x^T x_i + b^*) \qquad \text{式(3-9)}$$

2） 线性不可分的广义最优分类面

最优分类面是在线性可分的前提下讨论的，在大多数的实际应用中，并不都能满足线性可分性。在线性不可分的情况下，就是某些训练样本不能满足式(3-6)的条件，可以在条件中增加一个松弛项 $\Xi = (\xi_1, \xi_2, \ldots, \xi_l)^T, \xi_i = 0$ 调整约束条件，新的求最优分类面的公式变成：

$$\begin{aligned} &\underset{\omega,b,\Xi}{\text{Minimize}} \ \ \Phi(\omega,b,\Xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} \xi_i^k \\ &\text{subject to } y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i, \ \ \xi_i \geq 0, \ i = 1,...,l \end{aligned} \qquad \text{式(3-10)}$$

其中的参数 C > 0 被称为惩罚因子，它控制对错样本惩罚的程度，它决定了分类误差与类别间距之间的折中，折衷考虑最小错分样本和最大分类间隔就可以得到广义最优分类面，而所有非零的 ξ_i 决定了分类的误差。一般需要根据实验，发现合适的 C 值。

3） 支持向量机的核函数

若在原始空间中的简单超平面不能得到满意的分类效果，则必须以复杂的超平面作为分界面。对于空间 L 内非线性分类问题，可以通过一个非线性变换 $\phi(x)$ 将数据 x 从原空间 L 映射到一个高维特征空间 H，再在新空间 H 中建立最优分类面。这时的分类函数是：

$$f(x)=\mathrm{sgn}(\phi(x)^T\omega^*+b^*)=\mathrm{sgn}(\sum_{i=1}^{l}y_i\lambda_i^*\phi(x)^T\phi(x_i)+b^*) \quad \text{式(3-11)}$$

我们为了计算式(3-11)，并不需要知道$\phi(x)$的具体形式，因为根据 Hibert-Schmidt 定理，可以通过核函数$K(x,y)$间接得到该值，即

$$K(x,y)=\phi(x)^T\phi(y) \quad \text{式(3-12)}$$

于是，用核函数$K(x,y)$代替最优分类面中的点积x^Tx_i，就相当于把原始特征空间变换到了新的特征空间，最终的判断函数是

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}y_i\lambda_i^*K(x,x_i)+b^*) \quad \text{式(3-13)}$$

可见，非线性变换是通过定义适当的核函数实现的。这种核函数的变换处理，为支持向量机提供了极大的灵活性，使其有了更广泛的应用范围。常见的核函数类型有：

（1） 多项式核函数

$$K(x_i,x_j)=(\gamma(x_i^Tx_j)+\gamma) \quad \text{式(3-14)}$$

对应 SVM 是一个d阶多项式分类器；

（2） 径向基核函数（RBF）

$$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2+\gamma) \quad \text{式(3-15)}$$

对应 SVM 是一种径向基函数分类器；

（3） Sigmoid 核函数

$$K(x_i,x_j)=\tanh(\gamma(x_ix_j)+\gamma) \quad \text{式(3-16)}$$

对应 SVM 实现的是一个两层的感知器神经网络。

3.2 SVM 分类器的置信度评估算法

本节将详细介绍一类对余类（one-against-all，简称 OAA 方法）SVM 分类器和一类对一类（one-against-one，简称 OAO 方法）SVM 分类器的基本理论和置信度评估算法。

3.2.1 “一类对余类”方法

3.2.1.1 “一类对余类”方法理论基础[10][33]

SVM 多类分类方法最早使用的算法就是“一类对余类”方法。要得到多类分类器，通常的方法是构造一系列两类分类器，其中的每一个分类器都把其中的一类同余下的各类分划开，也就是把某一类别的样本当作一个类别，剩余其他所有类别构成的整体作为另一个类别，组成两类问题，然后据此推断待测样本 x 的归属。“一类对余类”方法是对于 k 分类问题构造 k 个 SVM 子分类器。在构造第 i 个 SVM 子分类器时，将属于第 i 类别的样本数据标记为正类，不属于 i 类别的样本数据都标记为负类。在决策过程中，对测试样本分别计算各个子分类器的决策函数值，并选取分类器函数值最大所对应的类别作为测试样本的预测类别。第 i 个 SVM 子分类器需要解决的最优化问题如下：

$$\begin{cases} \underset{\omega^i,b^i,\Xi^i}{Minimize}\ \Phi(\omega^i,b^i,\Xi^i)=\dfrac{1}{2}\left\|\omega^i\right\|^2+C\sum_{j=1}^{l}\xi_j^i \\ s.t.\ (\omega^i)^T\phi(x_j)+b^i\geq 1-\xi_j^i, if\ y_j=i \\ (\omega^i)^T\phi(x_j)+b^i\leq -1+\xi_j^i, if\ y_j\neq i \\ \xi_j^i\geq 0, j=1,2,\ldots,l \end{cases} \qquad \text{式(3-17)}$$

解决该式的最优化问题后，就可以得到 k 个决策函数：

$$\begin{cases} (\omega^1)^T\phi(x)+b^1 \\ \cdot \\ \cdot \\ \cdot \\ (\omega^k)^T\phi(x)+b^k \end{cases} \qquad \text{式(3-18)}$$

对于待测样本 x，将其输入这 k 个决策函数中，得到 k 个值，取得最大值的决策函数所对应的类别即为该测试样本所属类别。

3.2.1.2 一类对余类 SVM 分类器的置信度评估算法

凌萍，周春光等人[56]研究了一类对余类 SVM 多分类器的置信度评估。在训练阶段，生成 m 个一类对余类 SVM 自分类器，其决策函数分别为 f_1，f_2，…，f_m。在测试阶段，对待测数据 Q，生成各基本 SVM 决策函数 f_1，f_2，…，f_m 的置信度系数：λ_1，λ_2，…，λ_m。文中提出，各基本 SVM 的信用度取决于其自身的分类能力和在 Q 的邻域之上生成决策的一致程度。文中将各决策函数的置信度定义为：

$$\lambda_j=\exp(-\left\|\omega_j\right\|/\eta_j) \qquad \text{式(3-19)}$$

其中，$2/\|\omega_j\|$ 为决策函数自身分界面的间隔大小，η_j 为决策函数在以欧式测度建立的邻域内的分类准确率。定义

$$A1=\max\left\{f_j(Q)*\lambda_j\right\} \qquad \text{式(3-20)}$$

$$A2 = \text{second\_max}\left\{f_j(Q) * \lambda_j\right\} \qquad \text{式(3-21)}$$

并且，令 I 为 $A1$ 对应的类别，J 为 $A2$ 对应的类别。因此，若 Q 的决策风险值 $A1-A2$ 满足 $A1-A2 > e$，则将 Q 判断为 $A1$ 相对应的类 I，否则进行决策修正。

但是，该文提出的判断决策风险的方法并只适用于一类对余类 SVM 分类器。这种分类器的缺点是：每个分类器的训练都是将全部的样本作为训练样本，每个 SVM 的训练速度随着训练样本数量的增加急剧减慢。并且，该文中的置信度公式定义中的参数 η_j 为决策函数在以欧式测度建立的邻域内的决策正确率，此参数是由在线计算得出，引入了人工参与的过程。因此，本论文着手研究了另一种常用的 SVM 分类方法，即一类对一类 SVM 分类器的置信度评估算法。

3.2.2 “一类对一类”方法

3.2.2.1 “一类对一类”方法理论基础[10][33]

这种方法也是基于两类问题的分类方法。“一类对一类”分类器的具体分类方法是分别选取 2 个不同类别构成一个 SVM 子分类器，也就是每一类与其它各类别分别构成一个两类问题，这样共有 $k(k-1)/2$ 个 SVM 子分类器。在构造类别 i 和类别 j 的 SVM 子分类器时，在训练样本集选取属于类别 i、类别 j 的样本数据作为训练样本数据，并将属于类别 i 的样本标记为正，将属于类别 j 的样本标记为负。“一类对一类”方法需要解决的最优化问题如下：

$$\begin{cases} \underset{\omega^{ij},b^{ij},\xi^{ij}}{Minimize}\dfrac{1}{2}\left\|\omega^{ij}\right\|^2 + C\sum_t \xi_t^{ij} \\ s.t.\ (\omega^{ij})^T\phi(x_t) + b^{ij} \geq 1-\xi_t^{ij}, if\ y_t = i \\ (\omega^{ij})^T\phi(x_t) + b^{ij} \leq -1+\xi_t^{ij}, if\ y_t \neq i \\ \xi_t^{ij} \geq 0 \end{cases} \qquad \text{式(3-22)}$$

解决这一最优化问题后，也即用训练样本进行训练后就可以得到 $k(k-1)/2$ 个 SVM 子分类器。

在测试阶段，将测试样本对所有的 $k(k-1)/2$ 个 SVM 子分类器分别进行测试，得到各个识别结果，采用投票法来决定测试样本的类别。即经过 $k(k-1)/2$ 个 SVM 分类函数，累计出现最多次数的类别就是测试样本的最终预测类别。

本文中进行研究的 Libsvm 分类器就是这样的一类对一类的分类器。

3.2.2.2 一类对一类 SVM 分类器的置信度评估算法

3.2.1.2 节分析了已有的 SVM 分类器置信度评估算法存在的问题，本节就一类对一类 SVM 分类器的二分类问题提出了两组置信度评估公式（分别是置信度

评估公式 F1 和 F2），用以评估待测样本的分类结果，当给定一个拒识率时，接受高置信度样本的分类结果，拒识低置信度样本的分类结果。并且，对被拒识的这部分样本的分类结果，我们分别采用两种后处理方法（分别是拒识取反 OPP 和 KNN）进行处理。因此，本文共设计了 F1 - OPP，F1 - KNN，F2 - OPP，F2 - KNN 四种算法。

3.2.2.2.1 SVM 置信度评估及决策修正算法 F1 - OPP，F1 - KNN

李蓉等人[57]中提出，当待测样本到 SVM 分类面的距离大于一定的阈值时，则接受 SVM 分类结果。由于待测样本到 SVM 最优超平面的距离

$$d(x_i) = \mathrm{sgn}(\omega x + b)/|\omega| = \mathrm{sgn}(\sum_{j=1}^{l} y_j \alpha_j K(x_j x_i) + b)/|\omega| \qquad 式(3-23)$$

所以当待测样本到 SVM 最优超平面的距离大于给定的阈值，也就是当决策函数值大于一定的阈值时，我们接受 SVM 的分类结果。

我们认为，在 SVM 置信度评估公式 1 中，待测样本到 SVM 最优超平面的距离和置信度之间呈正比。即待测样本到分类面距离越大，该样本置信度相对越高；反之，待测样本到分类面距离越小，该样本的置信度相对越低。

根据上述分析，我们将置信度评估公式 1 定义如下：

$$f(x_i) = \exp(-1/|d(x_i)|) \qquad 式(3-24)$$

式中，$d(x_i)$为待测样本 x_i 到 SVM 分类面的距离。

根据上述分析，可以得到 SVM 置信度评估公式 1 所对应的算法框架设计，如图 3 - 2 所示：

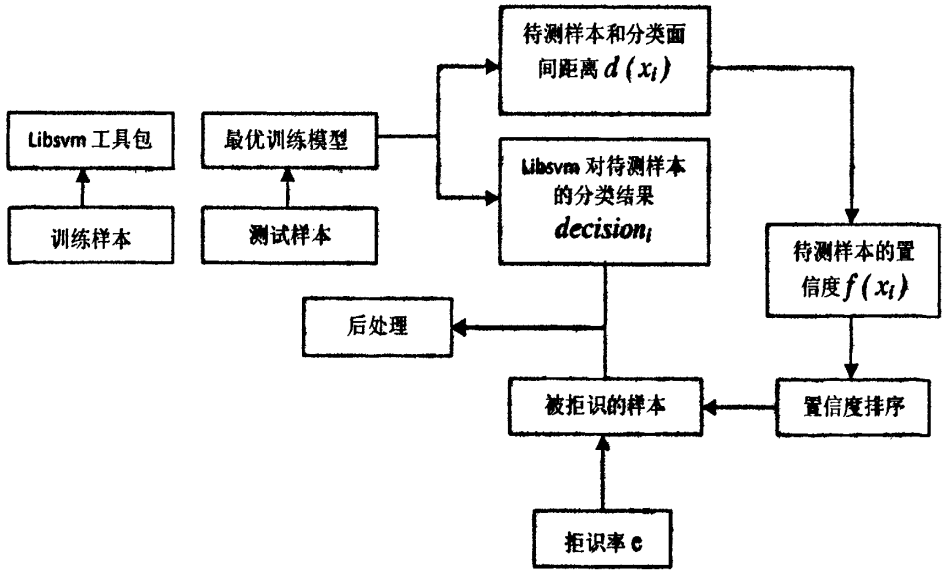

图 3-2 置信度评估及决策修正算法框架 1

在 Libsvm 训练阶段中，对训练样本，使用 Libsvm 工具包得出最优训练模型 *Model*，也即在 RBF 核函数下得出最优 C 参数和 g 参数。在 Libsvm 测试阶段中，用该模型预测待测样本，获取每个待测样本 x_i 到 SVM 最优分类面的距离 $d(x_i)$，以及 Libsvm 对每个测试样本 x_i 的分类决策结果 $decision_i$。根据置信度评估公式 1 计算出每个待测样本的置信度 $f(x_i)$。若测试样本数为 n，对所有待测样本的置信度 $f(x_1), f(x_2), f(x_3), \cdots, f(x_n)$ 由低到高进行排序，得出置信度排序后的序列 $g_1, g_2, g_3, \cdots, g_n$。若给定拒识率 e，应拒识排序前 $e*n$ 个待测样本的分类结果。也就是说，对于给定的拒识率 e，置信度阈值为 g_{e*n+1}。因此，应当拒识的样本为置信度小于阈值 g_{e*n+1} 的样本。

由于被拒识的为置信度低的样本，即 Libsvm 对这部分样本的分类结果 $decision_i$ 具有高风险。对这部分样本，我们设计了以下两种方法进行后处理：算法 F1 - OPP 和算法 F1 - KNN。1、算法 F1 - OPP：该算法将拒识的这部分样本的分类结果 $decision_i$ 先取反再输出。即经过 Libsvm 决策，分类结果 $decision_i$ 为正的样本，将其处理为负样本；反之，处理为正样本。由于被拒识的为低置信度的样本，即该部分样本的分类结果具有高风险，采取直接对这部分样本分类结果取反的方法，经过最终的结果统计分析，就能够验证我们是否较准确地定位了需拒识的样本，也就能验证定义的置信度评估公式 1 能否较准确的评估待测样本的置信度。2、算法 F1 - KNN：该算法将被拒识的这部分样本带入基于欧式距离的 KNN 分类器进行重新分类。

(1) 算法 F1 - OPP 具体描述如下：

输入：训练样本集 S，待测样本集 $T(x_1, x_2, x_3, \cdots, x_n)$，拒识率 e

输出：直接接受的这部分测试样本分类结果以及经过拒识取反后处理的剩下样本分类结果。

Step 1：$Model = Libsvm(S)$；

//对训练样本，使用 Libsvm 工具包得到最优训练模型 *Model*

Step 2：$i = 1$；

//初始化

Step 3：如果 $T \neq \Phi$，取 $x_i \in T$，如果 $T = \Phi$，转 Step 7；

//依次取待测样本集中的测试样本

Step 4：$Model(T) \rightarrow (d(x_i), decision_i)$；

//使用最优训练模型 *Model* 预测待测样本，获取待测样本 x_i 到 SVM 最优分类面的距离 $d(x_i)$，以及 Libsvm 对待测样本的分类结果 $decision_i$

Step 5：$d(x_i) \rightarrow f(x_i)$；

//根据置信度评估公式 1，计算待测样本的置信度

Step 6：$i = i + 1$，转 Step 3；

Step 7：Sort $(f(x_1), f(x_2), f(x_3), \cdots, f(x_n)) \rightarrow g_1, g_2, g_3, \cdots, g_n$；

//对所有待测样本的置信度从小到大进行排序

Step 8：$\theta = g_{e*n+1}$；

//给定拒识率 e，拒识置信度序列 $g_1, g_2, g_3, \cdots, g_n$ 中排序最小的 $e * n$ 个，也就是令置信度拒识阈值为 g_{e*n+1}

Step 9：$REJ = reject\,(f(x_i) < \theta)$即 $REJ = reject\,(g_1, g_2, g_3, \cdots, g_{e*n})$，$ACC = accept\,(f(x_i) = \theta)$；

//接受置信度大于阈值的相应样本，直接输出这些样本的分类结果，拒识剩余样本，放入拒识样本集

Step 10：$ACC\,(opp_i = -\,decision_i)$；

//对拒识样本集中样本的初始分类结果 $decision_i$ 取反并输出

Step 11：END。

//算法结束

(2) 算法 F1 - KNN 具体描述如下：

输入：训练样本集 S，待测样本集 $T(x_1, x_2, x_3, \cdots, x_n)$，拒识率 e

输出：直接接受的这部分样本的分类结果以及经过 KNN 后处理的剩下样本分类结果。

Step 1：$Model = Libsvm\,(S)$；

//对训练样本，使用 Libsvm 工具包得到最优训练模型 *Model*

Step 2：$i = 1$；

//初始化

Step 3：如果 $T \neq \Phi$，取 $x_i \in T$，如果 $T = \Phi$，转 Step 7；

//依次取待测样本集中的测试样本

Step 4：$Model\,(T) \rightarrow (d(x_i), decision_i)$；

//使用最优训练模型 *Model* 预测待测样本，获取待测样本 x_i 到 SVM 最优分类面的距离 $d(x_i)$，以及 Libsvm 对待测样本的分类结果 $decision_i$

Step 5：$d(x_i) \rightarrow f(x_i)$；

//根据置信度评估公式 1，计算待测样本的置信度

Step 6：$i = i + 1$，转 Step 3；

Step 7：Sort $(f(x_1), f(x_2), f(x_3), \cdots, f(x_n)) \rightarrow g_1, g_2, g_3, \cdots, g_n$；

//对所有待测样本分类结果的置信度从小到大进行排序

Step 8: $\theta = g_{e*n+1}$;

//给定拒识率 e，拒识置信度序列 $g_1, g_2, g_3, \cdots, g_n$ 中排序最小的 $e*n$ 个，也就是令置信度拒识阈值为 g_{e*n+1}

Step 9: $REJ = reject(f(x_i) < \theta)$即 $REJ = reject(g_1, g_2, g_3, \cdots, g_{e*n})$, $ACC = accept(f(x_i) = \theta)$;

//接受置信度大于阈值的相应样本，直接输出这些样本的分类结果，拒识剩余样本，放入拒识样本集

Step 10: $ACC(KNN(f(x_i) < \theta))$;

//将被拒识的样本带入 KNN 分类器重新分类并输出

Step 11: END。

//算法结束

3.2.2.2.2 SVM 置信度评估及决策修正算法 F2 - OPP，F2 - KNN

可以发现，置信度评估公式 1 定义的置信度只考虑了待测样本 x_i 到最优分类面的距离 $d(x_i)$这一个参数。通过观察分析，我们发现除了待测样本 x_i 到分类面距离这个参数外，还有一个可能对置信度有影响的参量，即待测样本邻域内的训练样本与其属于同一个类别的概率。具体来说，经过 Libsvm 判断决策，若待测样本 x_i 为正样本，而与该待测样本距离最近的 j 个训练样本中，正样本的个数越多，表明 x_i 实际为正的可能性越大，正样本的个数越少，表明 x_i 实际为正的可能性也越小。因此，待测样本周围 j 个训练样本点与其同属一类的概率 ρ_i 的大小，可以相对反应出 Libsvm 对该待测样本的分类正确的可能性大小，参数 ρ_i 应当做为置信度评估公式的一个参量，并且该参量同置信度之间呈正比。

根据上述分析，我们将置信度评估公式 2 定义如下：

$$f(x_i) = \exp(-1/(|d(x_i)| * \rho_i)) \qquad \text{式(3-25)}$$

式中，$d(x_i)$为待测样本 x_i 到 SVM 分类面距离，即

$$d(x_i) = \mathrm{sgn}(\omega x + b)/|\omega| = \mathrm{sgn}(\sum_{j=1}^{i} y_j \alpha_j K(x_j x_i) + b)/|\omega| \qquad \text{式(3-26)}$$

ρ_i 为待测样本周围 j 个训练样本点属于待测样本经 Libsvm 判断的分类结果 $decision_i$ 这一类的概率，即

$$\rho_i = j_s / j \qquad \text{式(3-27)}$$

其中，j 为样本近邻点个数，j_s 为待测样本经过 Libsvm 判断所属类别包含的近邻点的个数。

根据上述分析，可以得到 SVM 置信度评估公式 2 所对应的算法框架设计，

如图 3 - 3 所示：

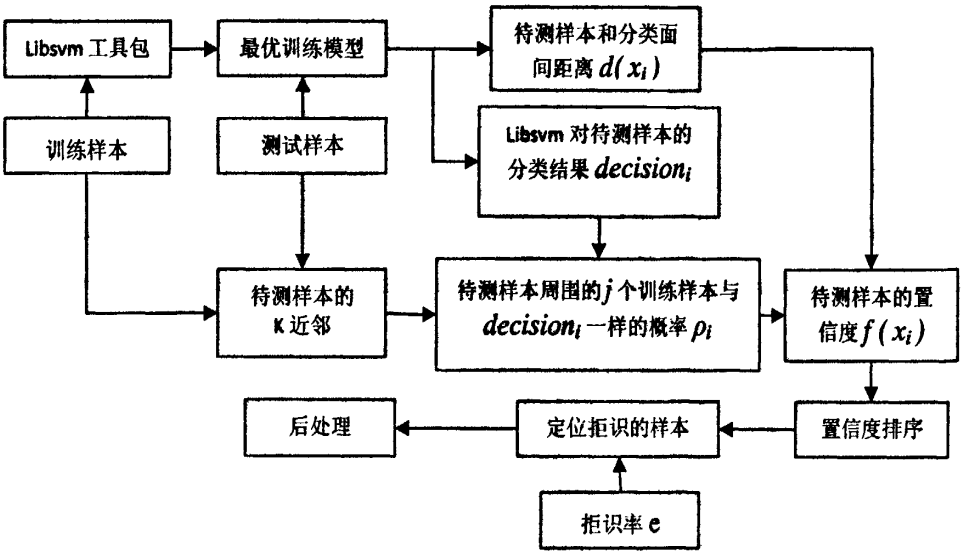

图 3-3　置信度评估及决策修正算法框架 2

在 Libsvm 训练阶段中，对训练样本，使用 Libsvm 工具包得到最优训练模型 *Model*，也就是在 RBF 核函数下得出最优 C 参数和 g 参数。在 Libsvm 测试阶段中，用该最优模型 *Model* 预测待测样本，获取每个待测样本 x_i 到 SVM 最优分类面的距离 $d(x_i)$，以及 Libsvm 对每个测试样本 x_i 的分类预测结果 $decision_i$。并且，在训练样本里找到 x_i 的 j 个近邻，计算出这 j 个近邻中和 x_i 的初始分类结果 $decision_i$ 一样的概率 ρ_i。结合 $d(x_i)$ 和 ρ_i 这两个参数，根据置信度评估公式 2，我们可以得出该待测样本 x_i 的置信度 $f(x_i)$。若测试样本数为 n，对所有待测样本的置信度 $f(x_1), f(x_2), f(x_3), \cdots, f(x_n)$ 由低到高进行排序，得出排序后的序列 $g_1, g_2, g_3, \cdots, g_n$。给定拒识率 e，则应拒识排序前 $e*n$ 个测试样本的分类结果。也就是，根据给定的拒识率 e，能够得出置信度阈值 g_{e*n+1}，进而接受置信度大于阈值的样本，拒识置信度小于阈值的样本。

对被拒识的这部分样本，我们采用两种方法进行后处理：算法 F2 - OPP 和算法 F2 - KNN。1、F2 - OPP 算法：该算法将被拒识的这部分样本的分类结果 $decision_i$ 取反再输出。即经过 Libsvm 预测，分类结果 $decision_i$ 为正的样本，将其处理为负样本；反之，处理为正样本。2、F2 - KNN 算法：该算法将被拒识的这部分样本带入基于欧式距离的 KNN 分类器重新进行分类。

(1) 算法 F2 - OPP 具体描述如下：

输入：训练样本集 S，待测样本集 T $(x_1, x_2, x_3, \cdots, x_n)$，拒识率 e

输出：接受的部分样本分类结果以及经过拒识后处理的剩余样本分类结果。

Step 1：*Model* = *Libsvm* (S)；

//对训练样本，使用 Libsvm 工具包得到最优训练模型 *Model*

Step 2：$i = 1$；

//初始化

Step 3：如果$T \neq \Phi$，取$x_i \in T$，如果$T = \Phi$，转 Step 9；

//依次取待测样本中的测试样本

Step 4：$Model(T) \rightarrow (d(x_i), decision_i)$；

//使用最优训练模型 *Model* 预测待测样本，获取每个待测样本x_i到 SVM 最优分类面的距离$d(x_i)$，以及 Libsvm 对每个待测样本的分类结果$decision_i$

Step 5：$(x_i, S) \rightarrow n_{i1}, n_{i2}, n_{i3}, \cdots, n_{ij}$

//对每个测试样本，在训练样本中找到其j个近邻

Step 6：$(decision_i, n_{i1}, n_{i2}, n_{i3}, \cdots, n_{ij}) \rightarrow \rho_i$；

//根据 Libsvm 对当前测试样本的分类结果，得出待测样本的j个近邻训练样本属于$decision_i$一类的概率

Step 7：$(d(x_i), \rho_i) \rightarrow f(x_i)$；

//根据$d(x_i)$以及ρ_i计算待测样本的置信度

Step 8：$i = i + 1$，转 Step 3；

Step 9：$\text{Sort}(f(x_1), f(x_2), f(x_3), \cdots, f(x_n)) \rightarrow g_1, g_2, g_3, \cdots, g_n$；

//对所有待测样本的置信度从小到大进行排序

Step 10：$\theta = g_{e*n+1}$；

//给定拒识率 e，应当拒识置信度序列$g_1, g_2, g_3, \cdots, g_n$中排序最前的$e*n$个，即置信度拒识阈值为$g_{e*n+1}$

Step 11：$REJ = reject(f(x_i) < \theta)$即$REJ = reject(g_1, g_2, g_3, \cdots, g_{e*n})$并且$ACC = accept(f(x_i) = \theta)$；

//接受置信度大于阈值的相应样本，直接输出，拒识其他样本，放入拒识样本集

Step 12：$ACC(opp_i = -decision_i)$；

//对拒识样本集中样本的初始分类结果取反并输出

Step 13：END。

//算法结束

(2) 算法 F2 - KNN 具体描述如下：

输入：训练样本集 S，待测样本集 $T(x_1, x_2, x_3, \cdots, x_n)$，拒识率 e

输出：接受的部分样本分类结果以及经过拒识后处理的剩余样本分类结果。

Step 1：$Model = Libsvm(S)$；

//对训练样本，使用 Libsvm 工具包得到最优训练模型 *Model*

Step 2：$i = 1$；

//初始化

Step 3：如果$T \neq \Phi$，取$x_i \in T$，如果$T = \Phi$，转 Step 9；

//依次取待测样本中的测试样本

Step 4：$Model(T) \rightarrow (d(x_i), decision_i)$；

//使用最优训练模型 *Model* 预测待测样本，获取每个待测样本x_i到 SVM 最优分类面的距离$d(x_i)$，以及 Libsvm 对每个待测样本的分类结果$decision_i$

Step 5：$(x_i, S) \rightarrow n_{i1}, n_{i2}, n_{i3}, \cdots, n_{ij}$

//对每个测试样本，在训练样本中找到其j个近邻

Step 6：$(decision_i, n_{i1}, n_{i2}, n_{i3}, \cdots, n_{ij}) \rightarrow \rho_i$；

//根据 Libsvm 对当前测试样本的分类结果，得出待测样本的j个近邻训练样本属于$decision_i$一类的概率

Step 7：$(d(x_i), \rho_i) \rightarrow f(x_i)$；

//根据$d(x_i)$以及ρ_i计算待测样本的置信度

Step 8：$i = i + 1$，转 Step 3；

Step 9：$\text{Sort}(f(x_1), f(x_2), f(x_3), \cdots, f(x_n)) \rightarrow g_1, g_2, g_3, \cdots, g_n$；

//对所有待测样本的置信度从小到大进行排序

Step 10：$\theta = g_{e*n+1}$；

//给定拒识率 e，应当拒识置信度序列$g_1, g_2, g_3, \cdots, g_n$中排序最前的$e*n$个，即置信度拒识阈值为$g_{e*n+1}$

Step 11：$REJ = reject(f(x_i) < \theta)$即$REJ = reject(g_1, g_2, g_3, \cdots, g_{e*n})$并且$ACC = accept(f(x_i) = \theta)$；

//接受置信度大于阈值的相应样本，直接输出，拒识其他样本，放入拒识样本集

Step 12：$ACC(KNN(f(x_i) < \theta))$；

//将拒识样本带入 KNN 分类器重新分类并输出

Step 13：END。

//算法结束

3.3 小结

本章首先介绍了在统计学习理论基础上发展起来的新的机器学习方法——支持向量机，接着对前人已研究的一类对余类 SVM 分类器的置信度评估方法进行了介绍，接着详细说明了本文设计研究的 SVM 分类器的置信度评估公式，并

设计了 F1 - OPP，F1 - KNN，F2 - OPP 及 F2 – KNN 四种置信度评估及决策修正算法。

我们在下一章中的实验中，首先将 Libsvm 分类器的分类结果同算法 F1 - OPP 和算法 F2 – OPP 的实验结果进行比对，即对两个不同的置信度公式均采用拒识取反后处理方法进行决策修正，找到更准确的置信度评估公式。再用选出的置信度评估公式给出待测样本的置信度，用两种不同的后处理方法得出修正后的测试样本的分类结果，将两组结果进行统计分析，最终在上述四种算法中获得最佳的置信度评估及决策修正算法。

第四章 实验与结果分析

4.1 SVM 分类器——Libsvm 工具包

本文研究的是一类对一类 SVM 分类器，Libsvm 就是这样的工具包。Libsvm 是台湾大学林智仁(Chih-Jen Lin)博士等开发设计的一个操作简单、易于使用、快速有效的通用 SVM 软件包，可以解决分类问题、回归问题以及分布估计等问题，提供了线性、多项式、径向基和 S 形函数四种常用的核函数供选择，可以有效地解决多类问题、交叉验证选择参数、对不平衡样本加权、多类问题的概率估计等。 LIBSVM 是一个开源的软件包，能够免费的从作者的个人主页 http://www.csie.ntu.edu.tw/~cjlin/ 处获得。他不仅提供了 LIBSVM 的 C++ 语言的算法源代码，还提供了 Python、Java、R、MATLAB、Perl、Ruby、LabVIEW 以及 C#.net 等各种语言的接口，可以方便的在 Windows 或 UNIX 平台下使用，也便于科研工作者根据自己的需要进行改进[58][59]。

4.2 实验数据

本实验中的实验语料是第一届COAE（中文倾向性分析评测）大赛的生语料。由于原始文档是无法直接运用分类算法直接处理的，因此对文档进行分类之前需要将文档表示为计算机能够处理的形式。我们对生语料经过人工标注，得到中立和客观共2341个样本。其中，标注过程是由三个研究生进行的，并且均得到一致结果。

对生语料经分词、去停用词后，计算词的卡方值。根据卡方值排序，选择前 1130 个词作为特征，以该词在文本中出现的次数为特征值，将文档转化为特征向量。其中，分词使用了张素香在 SIGHAN[2006]开发的汉语切分工具[60]。

我们选择人工标注得到的中立这一类样本作为正样本，相应的，客观这一类样本作为负样本。从样本集中随机抽取出 600 个正样本 600 个负样本作为训练集，余下的 1141 个样本做测试集。

4.3 实验设计

为了更全面的分析并验证前文提出的置信度评估公式及决策修正算法对Libsvm分类性能的影响，我们设计了七组实验。

第一组实验使用Libsvm工具包得到最优训练模型，用该模型判断待测样本，得出Libsvm分类的分类性能。

第二组实验采用置信度评估公式1提出的置信度评估算法F1-OPP，采用拒识取反的方法处理被拒识的数据，得出最终分类性能。

第三组实验采用置信度评估公式2定义的置信度评估算法，对每个待测样本，在训练样本集中找到其周围5个近邻点，得出这5个近邻和待测样本x_i的分类结果$decision_i$一样的概率ρ_i，计算置信度，并采用拒识取反的方法F2-OPP处理被拒识的数据，得出最终分类性能。

对比以上三组实验的结果，可以得出哪个置信度评估公式更好地评估了Libsvm的分类结果。

第四组实验为不同规模训练样本下的算法稳定性验证实验。该组实验在训练样本集里随机抽取一部分样本作为训练样本来预测测试集里的所有样本，并采用拒识取反的方法F2-OPP处理被拒识的数据，得出最终分类性能。

第五组、第六组实验为后处理方法比较实验，将算法F2-OPP和算法F2-KNN的性能进行比较，验证用KNN进行后处理，是否比拒识取反的后处理方法，更能提高分类性能。

在依据以上几组实验，确定了最佳算法后，第七组实验用该算法对整个样本集（2341个样本）进行了交叉验证。

4.3.1 置信度评估公式验证实验

前三组实验的训练样本为均为训练样本集中的全部1200个样本，测试样本均为测试样本集中的全部1141个样本。

第一组实验使用Libsvm工具包对训练样本得到的最优训练模型预测待测样本，得到每个测试样本的分类结果$decision_i$，得出Libsvm分类的分类性能。

第二组实验采用置信度评估公式1提出的置信度评估算法F1-OPP，采用拒识取反的方法处理被拒识的数据，得出最终分类性能。

第三组实验采用置信度评估公式2定义的置信度评估算法F2-OPP，对每个待测样本x_i，找到其周围5个近邻训练样本点，得出这5个近邻和待测样本x_i

的分类结果 $decision_i$ 一样的概率 ρ_i，计算置信度，并采用拒识取反的方法处理被拒识的数据，得出改进后的分类性能。

本文采用正样本准确率(*Prec_p*)、正样本召回率(*Recall_p*)、正样本F-measure (*F_p*)、负样本准确率(*Prec_n*)、负样本召回率(*Recall_n*)、负样本F-measure (*F_n*)、准确率 *P* 作为实验一、二、三中的评测指标，：

i. *tp* — true positive（真正，TP）被模型预测为正的正样本；

ii. *fp* — false positive（假正，FP）被模型预测为正的负样本；

iii. *tn* — true negative（真负，TN）被模型预测为负的负样本 ；

iv. *fn* — false negative（假负，FN）被模型预测为负的正样本；

v. 正样本准确率（ *Prec_p* ）

$$Prec\_p = tp / (tp + fp) \qquad 式(4-1)$$

vi.正样本召回率（ *Recall_p*)

$$Recall\_p = tp / (tp + fn) \qquad 式(4-2)$$

vii.正样本 F-measure（ *F_p* ）

$$F\_p = (2 * Prcp\_p * Recall\_p) / (Prec\_p + Recall\_p) \qquad 式(4-3)$$

viii. 负样本（ *Prec_n* ）

$$Prec\_n = tn / (tn + fn) \qquad 式(4-4)$$

ix. 负样本召回率（ *Recall_n* ）

$$Recall\_n = tn / (tn + fp) \qquad 式(4-5)$$

x. 负样本 F-measure（ *F_n* ）

$$F\_n = (2 * Prcp\_n * Recall\_n) / (Prec\_n + Recall\_n) \qquad 式(4-6)$$

xi.准确率（ *P* ）

$$P = (tp + tn) / (tp + tn + fp + fn) \qquad 式(4-7)$$

实验一、二、三的结果如表 4 - 1 所示。

表 4 - 1 第一、二、三组实验结果

(%)	拒识率	*Prec_p*	*Recall_p*	*F_p*	*Prec_n*	*Recall_n*	*F_n*	*P*
第一组	0%	99.7097	83.4751	90.873	69.469	98.7421	81.5584	87.9053
第二组	1%	99.564	83.2321	90.6684	69.5364	99.0566	81.7121	87.6424
	2%	99.562	82.8676	90.4509	69.0789	99.0566	81.3953	87.3795
	3%	99.5646	83.3536	90.7407	69.6903	99.0566	81.8182	87.73

	4%	99.5614	82.7461	90.3782	68.9278	99.0566	81.2903	**87.2918**
	5%	99.5562	81.774	89.7932	67.7419	99.0566	80.4598	**86.5907**
	6%	95.5529	81.1665	89.4244	67.0213	99.0566	79.9492	**86.1525**
	7%	99.5522	81.045	89.3503	66.879	99.0566	79.8479	**86.0649**
	8%	99.7	80.6804	89.1874	66.5263	99.3712	79.6974	**85.8896**
	9%	99.6983	80.3159	88.9637	66.1088	99.3712	79.397	**85.6266**
	10%	99.6947	79.3439	88.3627	65.0206	99.3712	78.607	**84.9255**
第三组	1%	99.7097	83.4751	90.873	69.469	98.7421	81.5584	**87.9053**
	2%	99.5683	84.0826	91.1726	70.6278	99.0566	82.4607	**88.2559**
	3%	99.5708	84.5687	91.4586	71.267	99.0566	82.8947	**88.6065**
	4%	99.5714	84.6902	91.5299	71.4286	99.0566	83.004	**88.6941**
	5%	99.5702	84.4471	91.3872	71.1061	99.0566	82.7858	**88.5188**
	6%	99.572	84.8117	91.601	71.5909	99.0566	83.1135	**88.7818**
	7%	99.5726	84.9332	91.6721	71.754	99.0566	83.2232	**88.8694**
	8%	99.5714	84.6902	91.5299	71.4286	99.0566	83.004	**88.6941**
	9%	99.5733	85.0547	91.7431	71.9178	99.0566	83.3333	**88.9571**
	10%	99.7131	84.4471	91.4474	71.1712	99.3711	82.9396	**88.6065**

我们将上表最后一列的准确率 P 进行对比，如下图 4 - 1 所示

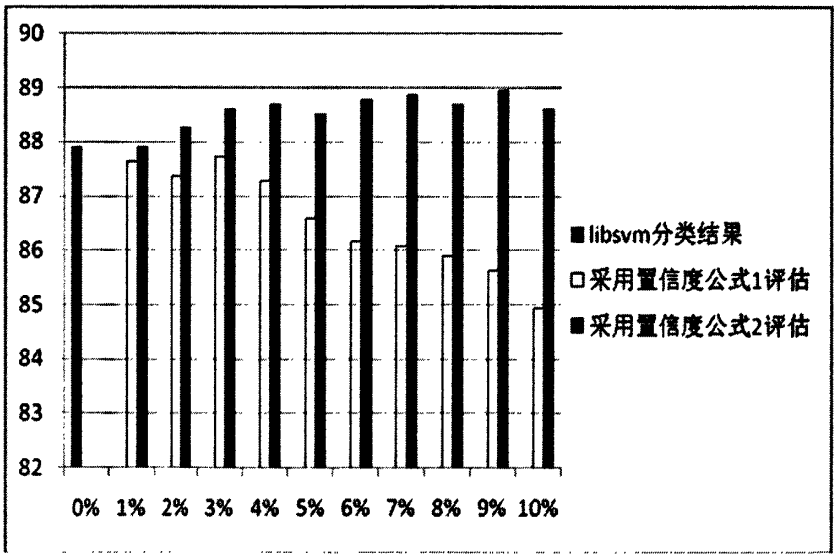

图 4-1 实验室一、二、三结果准确率对比

首先，对比第一组（Libsvm 分类结果）与第二组（采用算法 F1-OPP）实验数据可见，采用置信度公式 1 评估待测样本的置信度，无论拒识率取多大，对拒识样本的初始分类结果取反，最终分类结果均没能得到提高。这说明，置信度公式 1 并没能准确地评估待测样本的置信度，导致很多原本应当接受的样本被拒

识，应当被拒识的样本被直接接受，使得最终性能反而下降。

对比第一组与第三组（采用置信度公式2评估）实验结果可以发现，引入待测样本5个近邻与 $decision_i$ 同属一类的概率，在不同的拒识率下，实验结果均有1%左右的提高。这也证明了置信度公式2比置信度公式1能对Libsvm对待测样本分类结果的置信度进行更好的估计，也证明了对应当被拒识的样本的定位更加准确，对Libsvm最终分类性能的提高也更有帮助。

因此，在后续的实验中，我们均采用置信度公式2来评估待测样本分类结果的置信度。

4.3.2 不同规模训练样本下的算法稳定性验证实验

由于SVM在小样本分析下的分类性能良好，为了证明本文中的置信度评估算法在小样本训练的情况下也能够提高Libsvm的分类性能，第四组实验中训练数据的选取方法如下：在1200个训练样本集中进行随机抽取，分别抽取100个样本(其中50个正样本50个负样本)、200个样本(100个正样本100个负样本)、300个样本（150个正样本150个负样本)、……、1100个样本（550个正样本550个负样本）作为训练样本。并且，为了证明算法的稳定性，我们对训练样本随机抽取实验进行了5次，下面的实验结果表示的是每5次实验的统计平均值。也就是说，我们共进行了5*11=55次实验，最终统计的是每5次实验结果的统计平均值。其中待测样本分类结果的置信度采用置信度公式2进行评估，被拒识样本的处理采用直接取反的方法。

以下表4-2至表4-12中的数据，分别是每5次最终分类结果准确率P的统计平均值，即 $P_{ave} = (P_1 + P_2 + P_3 + P_4 + P_5) / 5$ 其中 P_1 是用第一次随机抽取的样本训练并最终得出的分类结果修正后的准确率，P_2 是用第二次随机抽取的样本训练并最终得出的分类结果修正后的准确率，以此类推。

并且，为了更直观的表现实验结果，分别对应作出图4-2至图4-12。

Group1 ：训练样本——50个正样本50个负样本

表4-2 100个训练样本下的F2-OPP结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	81.0167	80.8064	80.9115	81.0605	81.7529	81.8405

表4-2续

拒识率	6%	7%	8%	9%	10%
P_{ave}	81.9282	81.9107	82.2262	82.0333	81.8931

将表4-2结果表示成图4-2

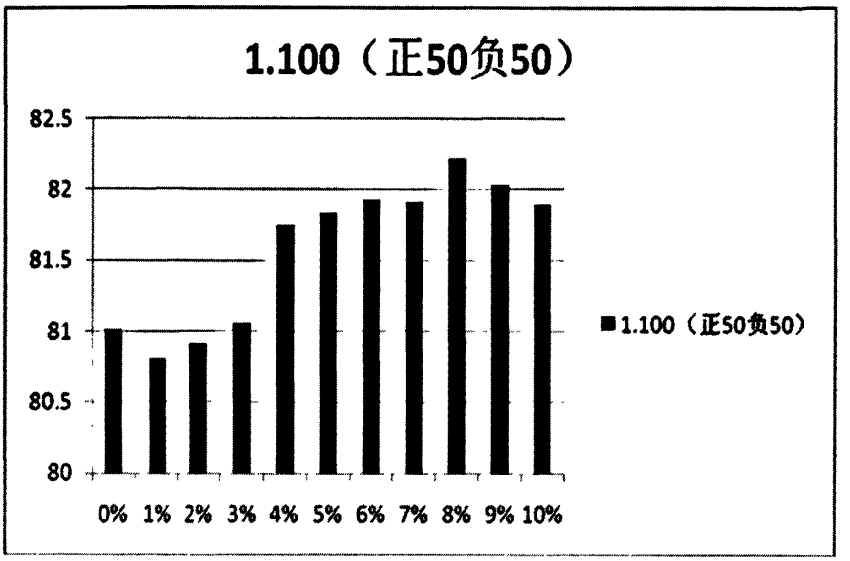

图 4－2 100 个训练样本下的 F2-OPP 结果

Group2 ： 训练样本——100 个正样本 100 个负样本

表 4－3 200 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	85.1358	85.2936	86.0473	86.4329	86.503	86.6608

表 4－3 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	86.5731	86.4154	86.2752	86.3628	85.9421

将表 4-3 结果表示成图 4-3

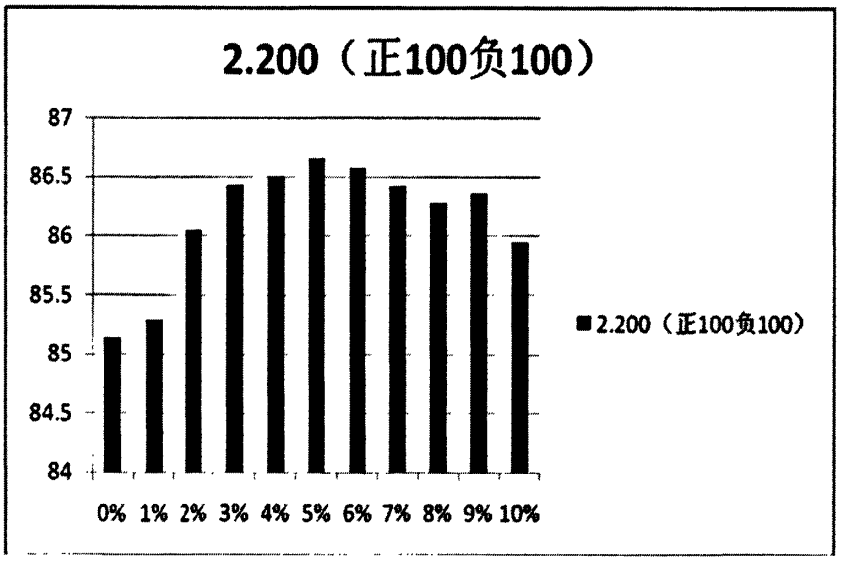

图 4－3 200 个训练样本下的 F2-OPP 结果

Group3 ： 训练样本——150 个正样本 150 个负样本

表 4－4 300 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	87.8177	87.6775	87.8177	87.9579	87.8528	88.326

表 4－4 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	88.2384	88.3961	88.5714	88.6941	88.4487

将表 4-4 结果表示成图 4-4

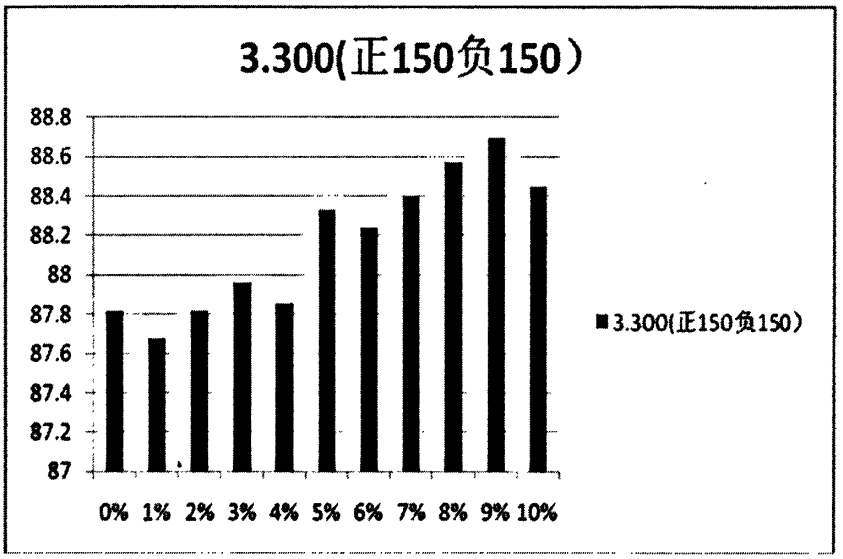

图 4－4 300 个训练样本下的 F2-OPP 结果

Group4：训练样本——200 个正样本 200 个负样本

表 4－5 400 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	85.8545	85.8545	86.2051	86.4505	86.6959	86.8186

表 4－5 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	87.0114	86.9237	86.8536	86.801	86.8361

将表 4-5 结果表示成图 4-5

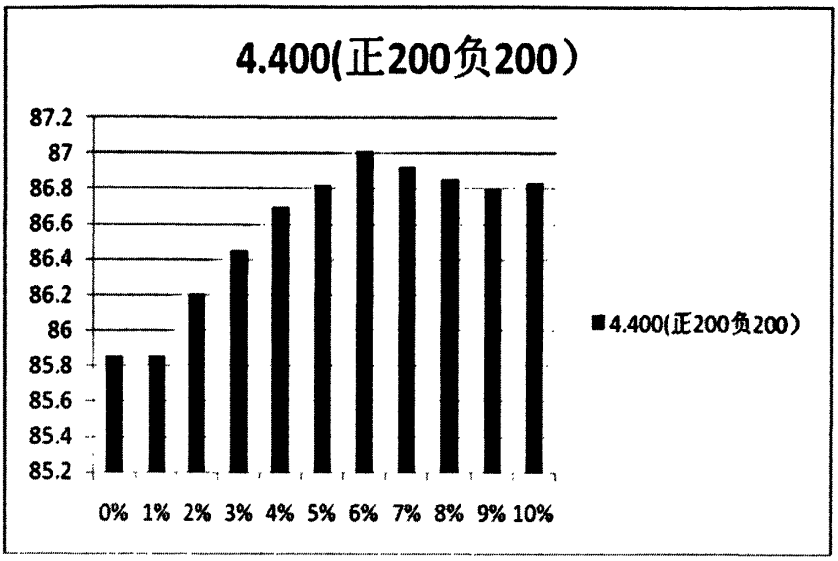

图 4－5 200 个训练样本下的 F2-OPP 结果

Group5：训练样本——250 个正样本 250 个负样本

表 4－6 500 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	86.9763	86.9763	86.766	87.397	87.5022	87.3795

表 4－6 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	87.3269	87.3093	87.2392	87.0114	86.7309

将表 4-6 结果表示成图 4-6

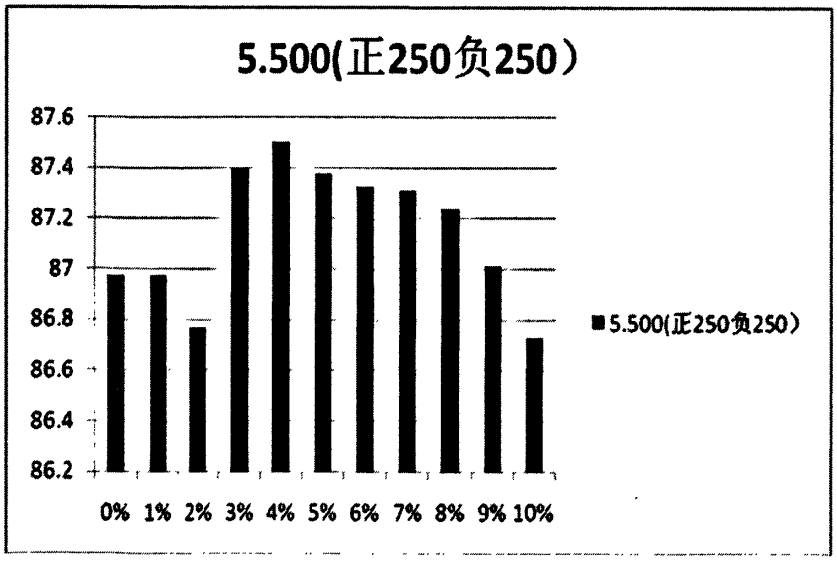

图 4－6 500 个训练样本下的 F2-OPP 结果

Group6：训练样本——300 个正样本 300 个负样本

表 4－7 600 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	85.3287	85.3287	85.6442	86.1	85.9773	86.0123

表 4－7 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	86.135	86.0474	85.837	85.8195	85.7494

将表 4-7 结果表示成图 4-7

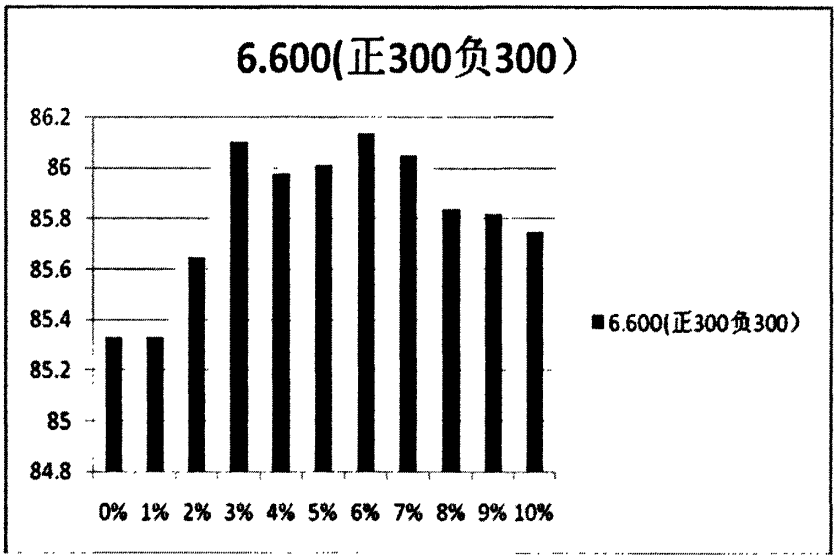

图 4－7 600 个训练样本下的 F2-OPP 结果

Group7：训练样本——350 个正样本 350 个负样本

表 4－8 700 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	85.4163	85.5039	85.837	86.3628	86.9413	86.8712

表 4－8 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	86.7835	86.9062	86.7309	86.7134	86.7835

将表 4-8 结果表示成图 4-8

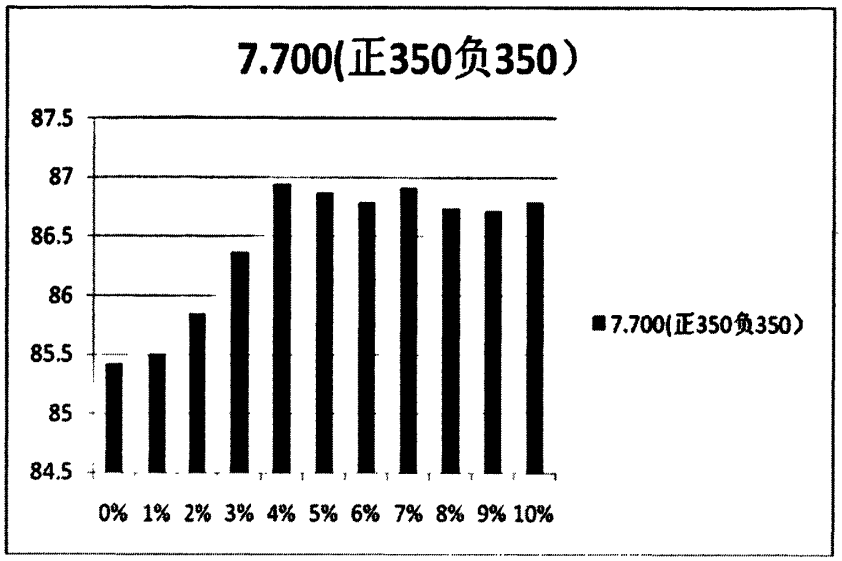

图 4－8 700 个训练样本下的 F2-OPP 结果

Group8：训练样本——400 个正样本 400 个负样本

表 4－9 800 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	88.1858	88.1858	88.1507	88.922	89.0096	89.15

表 4－9 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	89.0973	89.1148	88.9395	88.922	88.4663

将表 4-9 结果表示成图 4-9

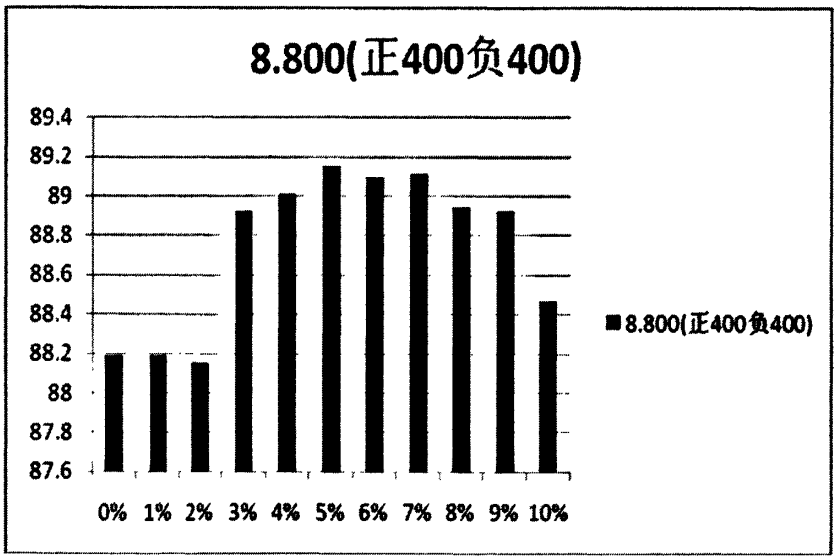

图 4－9 800 个训练样本下的 F2-OPP 结果

Group9：训练样本——450 个正样本 450 个负样本

表 4－10 900 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	82.3663	82.3663	82.3663	83.3128	83.7511	85.3987

表 4－10 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	85.5565	85.8545	85.7493	85.7318	85.9939

将表 4-10 结果表示成图 4-10

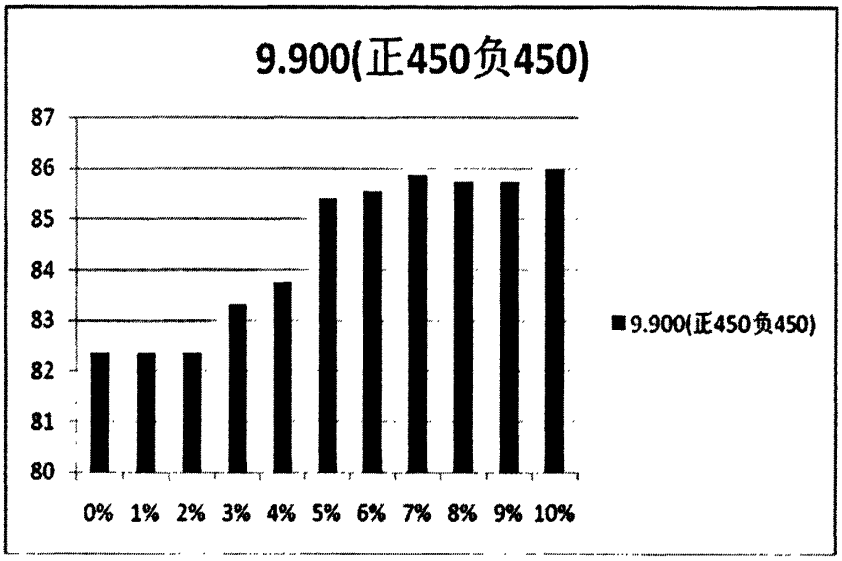

图 4－10 900 个训练样本下的 F2-OPP 结果

Group10：训练样本——500 个正样本 500 个负样本

表 4－11 1000 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	81.5776	81.5776	81.5776	81.7879	82.419	84.9606

表 4－11 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	84.9431	84.9255	85.1359	84.9781	84.7678

将表 4-11 结果表示成图 4-11

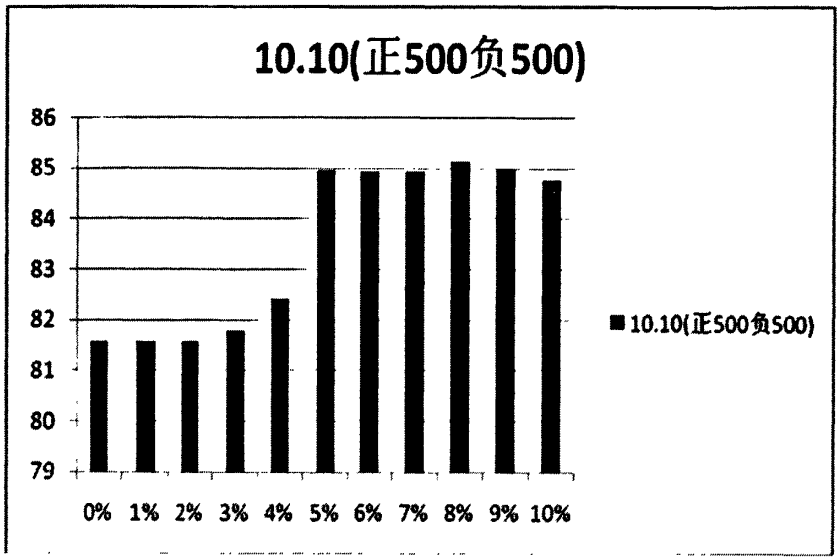

表 4－11 1000 个训练样本下的 F2-OPP 结果

Group11：训练样本——550 个正样本 550 个负样本

表 4－12 1100 个训练样本下的 F2-OPP 结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	84.61	84.61	84.61	86.1174	86.9413	87.2568

表 4－12 续

拒识率	6%	7%	8%	9%	10%
P_{ave}	87.1692	87.2568	87.2568	87.3094	87.2042

将表 4-12 结果表示成图 4-12

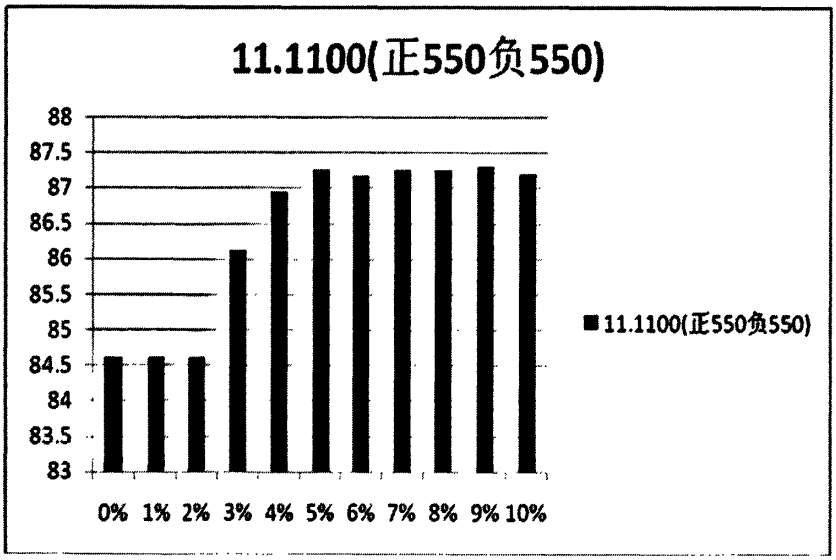

表 4－12 1100 个训练样本下的 F2-OPP 结果

观察以上各组实验结果，因为拒识率为 0%表示 Libsvm 的原始分类结果，

因此我们以拒识率 0%时的分类准确率作为基准，可以发现：当拒识率为 1% ~ 10%时，最终的待测样本的决策均能够得到不同程度的提高。这表明，算法 F2 - OPP 在不同规模训练样本下有着很好的稳定性，采用本算法对 Libsvm 初始预测结果进行修正均能够使分类准确率得到提高。并且，实验表明，在仅有少量样本参与训练的情况下，通过本文的拒识修正算法，我们提高 Libsvm 的初始预测结果，也能保证良好的分类准确率。

4.3.3 后处理方法比较

为了验证采用 KNN 分类器处理被拒识样本和对拒识的部分样本分类结果直接取反，这两种后处理方法哪个更有效，我们设计了第五组和第六组实验。这两组实验均采用了置信度评估公式 2 评估 Libsvm 的分类结果。置信度评估公式中的近邻点为 5，KNN 分类器中的 K 选取为 21。

第五组实验的训练语料为全部 1200 个训练样本，该组实验的准确率统计结果如表 4 - 13 和图 4-13 所示。

表 4 - 13 第五组实验结果

拒识率	0%	1%	2%	3%	4%	5%
F2 - OPP	87.9053	87.9053	88.2559	88.6065	88.6941	88.5188
F2 - KNN	87.9053	87.9053	88.3436	88.4329	89.3076	89.6582

表 4 - 13 续

拒识率	6%	7%	8%	9%	10%
F2 - OPP	88.7818	88.8694	88.6921	88.9571	88.6065
F2 - KNN	90.0964	90.447	90.6223	90.6860	90.6793

将表 4-13 结果表示成图 4-13

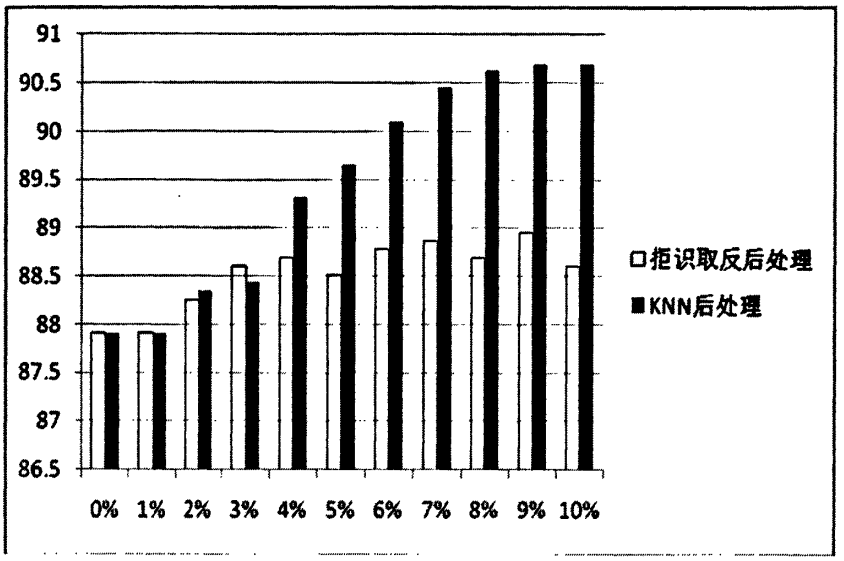

图 4 - 13 第五组实验结果

结果表明，对被拒识的这部分样本用 KNN 进行重新分类后的最终分类性能有了进一步的提高，比对被拒识的这部分样本判决类别直接取反的分类准确率更佳。

为了进一步验证算法 F2 - KNN 是否比算法 F2 - OPP 性能更佳，我们又进行了第六组实验。在第四组实验中，正 450 负 450 的五组训练样本经过算法 F2 - OPP 最终的分类准确率平均值得到了最大提高。因此，在第六组实验中，我们对同样的训练样本采用 F2 - KNN 算法处理数据。若待测样本的置信度小于给定拒识率相对应的置信度阈值，则带入 KNN 分类器重新分类。表 4 -14 和图 4 - 14 中统计的是 5 次实验中最终分类准确率的统计平均值。

表 4 - 14 第六组实验结果

拒识率	0%	1%	2%	3%	4%	5%
F2 - OPP	82.3663	82.3663	82.3663	83.3128	83.7511	85.3987
F2 - KNN	82.3663	82.3663	82.3663	83.2077	83.5232	85.539

表 4 – 14 续

拒识率	6%	7%	8%	9%	10%
F2 - OPP	85.5565	85.8545	85.7493	85.7318	85.9939
F2 - KNN	85.6792	86.3103	86.6082	86.8186	86.7649

将表 4-14 结果表示成图 4-14

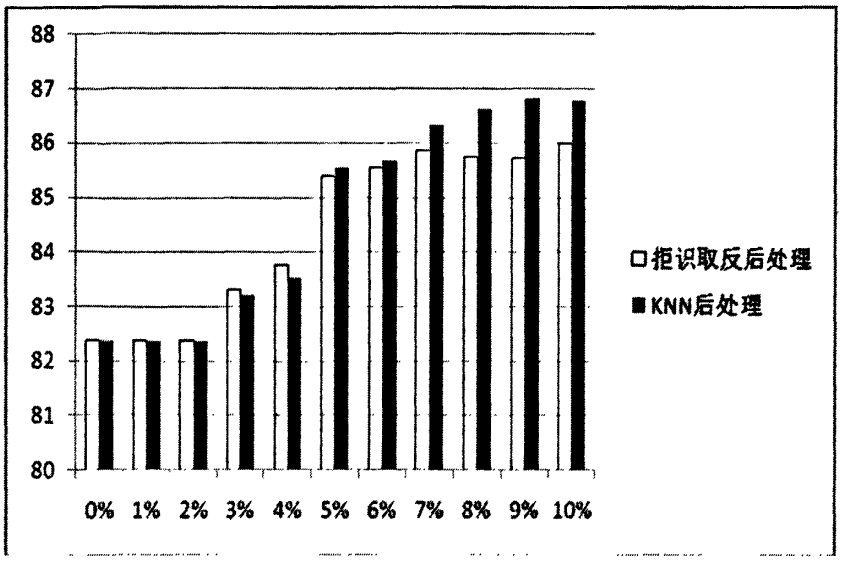

图 4 - 14 第六组实验结果

实验结果表明，对被拒识的这部分样本用 KNN 进行重新分类后的最终分类性能的确能够对系统识别准确率带来更大的提高，因此，算法 F2-KNN 更佳。另外，拒识率在 5%-10%的范围时，算法 F2-KNN 较大幅度地提高了 Libsvm 的分类性能，并且拒识率为 8%以上时，提高幅度变缓。

4.3.4 最终算法的交叉验证实验

在这组实验里，我们对算法 F2 - KNN 进行五折交叉验证：我们先将 2341 个训练样本随机划分为样本集 A（包含 468 个样本），样本集 B（包含 468 个样本），样本集 C（包含 468 个样本），样本集 D（包含 468 个样本），样本集 E（包含 469 个样本） 这 5 份。然后我们分别将其中的四份作为训练样本，用算法 F2 - KNN 预测剩下一份样本，即：1、样本集 A+B+C+D 作为训练样本，样本集 E 作为待测样本；2、样本集 A+B+C+E 作为训练样本，样本集 D 作为待测样本；3、样本集 A+B+ D+E 作为训练样本，样本集 C 作为待测样本；4、样本集 A +C+D+E 作为训练样本，样本集 B 作为待测样本；5、样本集 B+C+D+E 作为训练样本，样本集 A 作为待测样本。

将五组实验结果的平均值统计如表 4 - 15 和图 4 - 15 所示：

表 4 - 15 第七组实验结果

拒识率	0%	1%	2%	3%	4%	5%
P_{ave}	91.6844	91.6844	91.6844	92.094	92.5214	92.6316

表 4 – 15 续

拒识率	6%	7%	8%	9%	10%

P_{ave}	92.8032	92.8974	92.8124	92.8264	92.4251

将表 4-15 结果表示成图 4-15

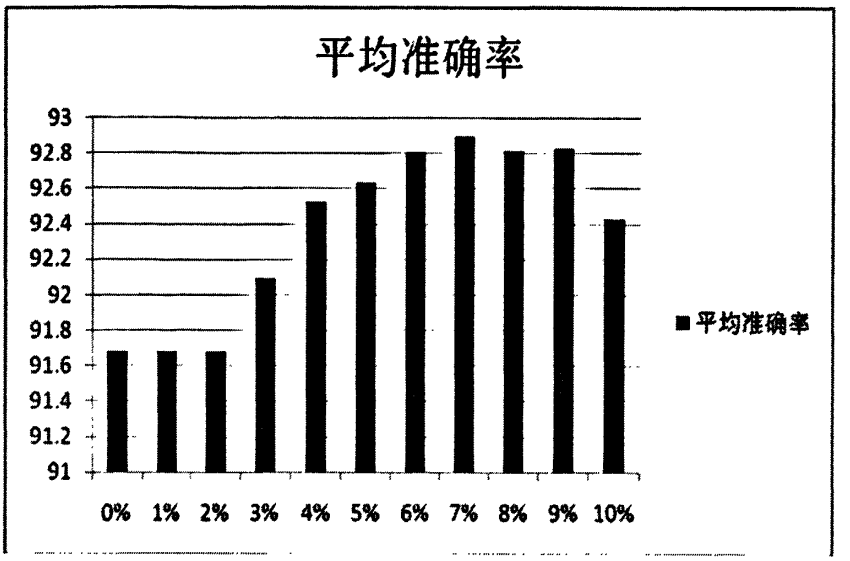

图 4 - 15 第七组实验结果

观察实验结果可以发现，即使是在 Libsvm 直接分类性能（91.6844%）已经非常好的情况下，采用本文提出的算法 F2 - KNN 对测试样本的分类结果进行修正，仍然能够进一步提高分类性能，并且当拒识率在 4%-10%的范围时，算法 F2-KNN 较大幅度地提高了 Libsvm 的分类性能，并且当拒识率为 8%以上时，性能提高幅度变缓，甚至下降。

4.4 小结

本章对前文提出的四种 SVM 置信度评估及修正算法 F1 - OPP、F1 - KNN、F2 - OPP 和 F2 - KNN，通过实验进行了对比验证。

实验结果表明，本文设计的置信度评估公式能够较为准确地评估待测样本的置信度，本文的拒识算法也能较好地提高 SVM 的分类结果。首先，置信度评估公式 2 能够更好地评估待测样本的置信度，将所有测试样本置信度排序后，给定拒识率，拒识置信度最小的样本。对被拒识的这部分样本，我们首先采用拒识部分取反的处理方法，但是这样并不能使分类准确率达到最佳。当我们在 Libsvm 后集成一个 KNN 分类器时，最终的分类性能得到了更大的提到。因此，本文提出的 F2 - KNN 算法为最终确定的最佳算法。并且，经过本论文的实验及分析，可以得出，拒识率的最佳取值范围为 5% - 8%。拒识率在此范围内，能够保证 Libsvm 的分类结果得到最大程度的提高。

第五章 总结与展望

5.1 研究工作小结

网络信息爆炸式的增长，促进了数据挖掘的应用，而分类技术作为数据挖掘技术的一个重要研究方面，受到研究者们的广泛关注。SVM作为最重要的分类技术之一，在许多实际分类应用中都表现了良好的性能。

本文对 Libsvm 这种一类对一类 SVM 分类器的二分类问题提出了置信度评估及决策修正算法，在该算法中，我们利用直接得到的观察量来反映识别结果之间的相对可靠性。在分类器预测阶段获取待测样本和最优分类超平面的距离，并且计算待测样本的 j 个近邻训练样本与待测样本经 Libsvm 判断的初始分类结果同属一类的概率。对于给定的拒识率，该算法拒识并修正置信度小于相应置信度阈值的样本分类结果。实验结果证明此置信度评估及修正算法能够很好地提高分类中常用的 Libsvm 分类器的性能，并且该算法具有相当的稳定性。

5.2 研究展望

本文提出的算法 F2-KNN 很好地提高了 SVM 二分类的分类性能，为 SVM 多分类时的置信度评估及修正问题的研究打下了良好的基础。虽然本文的实验数据选用的是文档，是在自然语言处理领域进行的应用，但是本文的算法适用于所有的 SVM 二分类问题。

本文的算法还有一些可以进一步细化和进一步研究的地方，可以从以下角度考虑：

进一步明确置信度评估公式里的参数 ρ_i——待测样本周围 j 个训练样本点属于待测样本经 Libsvm 判断的分类结果 $decision_i$ 这一类的概率。本文在计算该参数时，待测样本的近邻点数选择为 5，未综合考虑到训练样本集的数据分布情况等等。

再如拒识后处理中的 KNN 算法中，本文直接将 K 参数选取为 21，同样可以综合考虑训练样本集的数据分布情况来进行选取。

另外，拒识后处理中的 KNN 算法是基于欧式距离的，还可以采用多种形式考量 K 近邻相似值的大小，有待进一步研究比较。

同时，本文的实验数据是通过文本的向量表示的方法获得的，因此文本处理技术的成熟程度也影响了本方法的性能，但是我们在现有的条件下开展了本论文研究的工作，并且取得了不错的结果。随着文本处理技术的不断进步，本文所提出的方法在该领域的应用相信会获得更好的效果，在其他的领域也一定能表现出良好的性能。

参考文献

[1] 王继成,潘金贵,张福炎.Web 文本挖掘技术研究.计算机研究与发展.2005, 37(5):513-520.

[2] 赵辉. 基于 SVM 的数据挖掘分类技术研究[D].西安电子科技大学. 2008.

[3] Jiawei Han, Micheline Kamber 著. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2001 年 8 月.

[4] 史忠植著,知识发现[M]. 北京: 清华大学出版社, 2004 年.

[5] 陈海霞. 面向数据挖掘的分类器集成研究[D]. 吉林大学. 2006.

[6] 梁勇林. 基于多分类器融合的数据挖掘分类算法研究与应用[D]. 重庆大学. 2007.

[7] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995(20): 273-293.

[8] Vapnik 著.张学工译. 统计学习理论的本质[M]. 北京: 清华大学出版社, 2000.

[9] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社, 1999.

[10] 王静. SVM 在参数选择上的优化[J]. 兰州理工大学. 2008.

[11] Burges C, Schölkopf B. Improving the accuracy and speed of support vector learning machines. In Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA.1997:375-381.

[12] Platt J C. Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods-Support vector learning. Cambridge, MA: MIT Press, 1999.185-208

[13] J.A.K. Suykens, J. Vandewalle. Least squares support vector machine classifiers. Neural processing Letters, 1999, 9(3):293-300

[14] Osuna E, Freund R. Training Support Vector Machines: an application to face detection [A]. In: Proc of Computer Vision and Pattern Recognition[C]. San Juan，Puerto Rico，IEEE Computer Soc，1997:130-136

[15] Schwenker F. Hierarchical Support Vector Machines for Multi-class Pattern Recognition. In: Proceedings of the 4th International Conference on knowledge-based Intelligent Engineering Systems & Allied Technologies, Brighton UK. 2000:45-98

[16] 忻栋, 杨莹春, 吴朝晖.基于 SMV-HMM 混合模型的说话人确认[J].计算机辅助设计与图形学学报.2002,14(11):1080-1082

[17] W.M.Campbell. A SVM/HMM system for speaker recognition. In: Proceedings

of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003:209-212

[18] Yao Y, Gian L, Massimiliano P, et al. Combining flat and structured representations for fingerprint classification with recursive neural networks and support vector machines[J]. Pattern Recognition, 2003, 36(2):397-406

[19] Joachims T. Text categorization with support vector machines. Technical Report, LS VIII No.23, University of Dortmund, 1997

[20] 祝磊. 基于 SVM 技术的文本分类研究[J].软件导刊. 2006,(23):26-28.

[21] 马莉. 基于 SVM 的垃圾邮件过滤的研究[D].山东大学.2005.

[22] 王朝勇. 支持向量机若干算法研究及应用[D].吉林大学.2008.

[23] 张磊,林福宗,张钹.基于支持向量机的相关反馈图像检索算法[J].清华大学学报.2002, 42(1):80-83

[24] 肖靓.基于支持向量机的图像分类研究[D]. 同济大学. 2006.

[25] 付岩,王耀威,王伟强等. SVM 用于基于内容的自然图像分类和检索[J].计算机学报.2003, 26(10):1261-1265..

[26] 百度百科 http://baike.baidu.com/view/434404.htm

[27] 林晓帆, 丁晓青, 吴佑寿. 最近邻分类器置信度估计的理论分析[J].科学通报.1998,43(3):322-325.

[28] Cheng-Lin Liu, Masaki Nakagawa. Precise Candidate Selection for Large Character Set Recognition by Confidence Evaluation. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.22, no.6, 2000.

[29] 张丽. 基于多分类器动态组合的手写数字识别[D].南京理工大学.2003.

[30] 曹振华, 赵平. 概率论与数理统计(第一版)[M].东南大学出版社, 2004 年.

[31] 百度百科 http://baike.baidu.com/view/336751.htm?fr=ala0_1

[32] 王利民. 贝叶斯学习理论中若干问题的研究[D].吉林大学.2005.

[33] 叶志刚. SVM 在文本分类中的应用[D].哈尔滨工程大学.2006

[34] 郭亚琴. 分类器设计及组合技术研究[D].扬州大学.2007

[35] 林晓帆, 丁晓青, 吴佑寿等. 字符识别的置信度分析[J]. 清华大学学报.1998, 38(9):47-50.

[36] Smith S J. Handwritten character classification using nearest neighbor in large database. IEEE Trans PAMI, 1994, 16(9): 915-919

[37] Goudail F, Lange E, Iwamoto T, et al. Face recognition system using local autocorrelations and multiscale integration. IEEE Trans PAMI, 1996, 18(10): 1024-1028.

[38] 于一. K_近邻法的文本分类算法分析与改进[J].火力与指挥控制.2008, 33(4):143-145.

[39] 古平. 基于贝叶斯模型的文档分类及相关技术研究[D].重庆大学.2006.

[40] Yang Yiming & Liu Xin. A re-examination of text categorization methods. In: Proceeding of the 22th annual international ACM SIGIR conference on research and development in information retrieval.USA: ACM Press, 1999:42-49

[41] 薛磊, 杨晓敏, 吴炜等.一种基于 KNN 与改进 SVM 的车牌字符识别算法[J]. 四川大学学报.2006,43(5):1031-1036

[42] 李淑鹏. 基于神经网络的文本自动分类系统的研究[D].武汉理工大学.2008.

[43] 王志玲. 基于神经网络的文本自动分类系统研究[D].山东理工大学.2007.

[44] 全宏亮. 多层前向神经网络的结构辩识和改进算法[D].武汉科技大学.2002.

[45] 夏菁. 多层前向神经网络推广性研究及其应用[D].西北工业大学.2003.

[46] DE Rumelhart, JL McClelland. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge Bradford Books, MIT Press, 1986.

[47] Richard M D, Lippmann R P. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation, 1991, 3(4): 461-483

[48] 张岩. 基于语义角色的句子语义倾向判断[D].北京邮电大学. 2008.

[49] 王琪. 基于 SVM 的 Web 文本分类研究[D].上海海事大学. 2007.

[50] 陈平. 基于 SVM 的中文文本分类相关算法的研究与实现[D].西北大学.2008.

[51] 熊浩勇. 基于 SVM 的中文文本分类算法研究与实现[D].武汉理工大学.2008.

[52] 刘清. 基于 SVM 的网络文本分类问题研究与应用[D].南昌大学.2007.

[53] 李小英. 基于支持向量机的分类算法研究[D].东北电力大学.2008.

[54] Nello Cristianini, John Shawe-Taylor 著.李国正, 王猛, 曾华军译.支持向量机导论[M].北京:电子工业出版社,2004.

[55] Sergios Theodoridis, Konstantinos Koutroumbas 著.李晶皎等译.模式识别[M]. 北京:电子工业出版社, 2004.

[56] 凌萍,周春光.SVM 置信度在线评估以及决策改进[J].Journal of Frontiers of Computer Science and Technology, 2008,(2):192-197

[57] 李蓉,叶世伟,史忠植.SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J].电子学报,2002,(5):745-748

[58] 百度百科 http://baike.baidu.com/view/598089.htm?fr=ala0_1

[59] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a Library for Support Vector Machines Department of Computer Science, National Taiwan University, Taipei 106, Taiwan (http://www.csie.ntu.edu.tw/~cjlin)

[60] Suxiang Zhang, Ying Qin, JuanWen and Xiaojie Wang. "Word Seglnentation and Named Entity Recognition for SIGHAN Bakeoff3", Fifth SIGHAN Workshop on Chinese Language Processing, 2006.

致谢

在我攻读研究生硕士学位即将完成之际，回想我硕士期间的学习和生活，衷心地想对每一位帮助过我的人表示感谢！

首先衷心的感谢我的导师——何华灿教授。何老师严谨的治学态度、对科学认真钻研的精神等诸多品格，都深深地影响了我。

其次感谢我的指导老师——周延泉副教授，周老师对我的论文研究工作进行了全程的悉心指导。在我进行论文研究的过程中，周老师帮助我拓宽研究思路，帮助我改进我的研究方法。在我的学习、生活遇到困难的时候，周老师都给予我极大的支持和鼓励。周老师积极乐观的生活态度和处事方法始终让我受益匪浅，在北京邮电大学智能科学技术研究中心的学习生活将是我一生中难忘的时光。

在论文的完成过程中，王小捷老师从百忙中抽出时间，给予了悉心的指导，指出论文研究过程中的不足，在此，我向王小捷老师表示深深的谢意。

感谢智能科学技术研究中心的老师们，感谢钟义信教授在中期答辩过程中给我提出的意见和建议，李蕾老师、谭咏梅老师、李瑞凡老师等等都在平时的工作和学习中给予我很多的指导和帮助。老师们的言传身教，使我学习到了很多的知识，得到了很多的启发。

感谢课题组的毛昱师姐、李荣军师兄、潘文斌、赵文婧、郭叶、张予焓、王思宽、颜廷义、张博、张斌、旷远、张玉杰等同学，以及已经毕业的胡英飞师姐、马俊杰师兄、万鑫师兄等，在课题讨论过程中，大家的经验分享及大家对我的建议对我完成本论文都是很有帮助很有价值的。特别是毛昱师姐，每次同师姐的讨论都能为我带来启示，给了我很多帮助。

特别要感谢我的家人，感谢豆包，你们是我永远的动力。感谢你们为我付出的一切，感谢你们的支持和关爱。愿未来的日子里，你们和我共同幸福、共同快乐！

最后，感谢论文评审委员会的老师们百忙之中对我论文的悉心指正！

赵行

2010 年 1 月

攻读硕士学位期间发表的论文

[1] Xing ZHAO, Yanquan ZHOU, Huacan HE. Researches on Algorithm for Confidence Evaluation and Decision Modification of SVM. NLP-KE 2009(EI 检索)