兰州大学

硕士学位论文

基于主成分分析和支持向量机的组合判别分析方法研究

姓名：王惠婷

申请学位级别：硕士

专业：数学 应用数学

指导教师：王建州

20090501

摘 要

传统的判别分析是判别样品所属类型的一种统计方法，其应用之广可与回归分析媲美。常用的方法有距离判别法、Fisher判别法、Bayes判别法和逐步判别法。随着科学技术的飞速发展，关于人工智能的分类问题（模式识别，判别分析）的研究，已有大量的分类算法问世，例如：遗传算法、文本分类算法、贝叶斯分类算法、SVM（Support Vector Machina）分类算法、指纹分类算法等。但分类问题从理论上讲是一个复杂的函数延拓问题，不存在一种最优的分类算法适用于各种不同的情况，因此不存在一种判别分析方法适用于各种不同的情况，故至今仍有许多的判别分析方法陆续出现，我们也有必要对其进行不断的研究。

本文针对支持向量机分类的核函数选择问题，提出了新的方法—基于粒子群优化算法的组合核函数分类方法，该方法克服了单一核函数不能够精确描述判别结果的局限性。本文重点是解决某种指标繁多事物的分类判别问题，提出了基于主成分分析和支持向量机的组合判别分析方法，来对其事物所属种类进行判别分析，本文还给出了其数据分析的操作步骤。通过这种组合的判别分析方法，简化了运算，克服了线性方法的缺点，从而提高判断的准确率。最后应用于沙尘暴预警的研究，实验证明，这种组合判别方法是有效的。

关键词: 判别分析, 主成分分析, 支持向量机, 分类, 粒子群优化算法, 组合判别分析, 沙尘暴预警

Abstract

The traditional discriminant analysis is one statistical method for discriminating the sample respective type,widespread application to be on a par with regression analysis.Commonly used method have distance-distinguish method,fisher discriminant ,bayes discriminant ,stepwise discrimination for it.The rapid advancement in scientific and technological development,about artificial intelligence classified question (pattern recognition, discriminant analysis) research,it had the massive sorting algorithm to be published. For example,genetic algorithm,text classification algorithm,bayesian classification algorithm,SVM(support vector machine)for classification,fingerprint classification algorithm and so on. Because the classified question is the complex extension problem of function to explain theoretically,it does not have one kind of most superior sorting algorithm to be suitable for each kind of different situation.We have the necessity to its unceasing research.

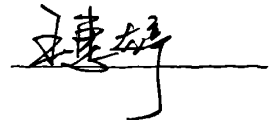
In order to solve the problem about choice of kernel functions in support vector machines for classification, in this paper we present a new approach—the combination kernel functions method for classification based on particle swarm optimization.This method overcame the limitations of the single nuclear function which describes inaccuracy discriminant result.In this paper,important aim at discriminant method for the thing having many indexs which have the harassment function to judge the result,we proposed based on the principal components analysis and the support vector machines combination discriminant analysis method for this kind of thing,and this article gives its sequence of operation.Through this method,both can simplify the operation,and overcome the flaw of linear methods,thus enhancement judgment rate of accuracy.Finally the application in the sand storm early warning's research,the experiment proved that the combination discriminant analysis method is effective.

Key words: discriminant analysis, principal components analysis, support vector machine, classification, particle swarm optimization algorithm, combination discriminant analysis

原创性声明

本人郑重声明：本人所呈交的学位论文，是在导师的指导下独立进行研究所取得的成果. 学位论文中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处. 除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果. 对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明.

本声明的法律责任由本人承担.

论文作者签名：  日期：2009.5.29

关于学位论文使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学. 本人完全了解兰州大学有关保存、使用学位论文的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本学位论文. 本人离校后发表、使用学位论文或与该论文直接相关的学术论文或成果时，第一署名单位仍然为兰州大学.

保密论文在解密后应遵守此规定.

论文作者签名：_____ 导师签名：王建州 日期：2009.5.29

第一章　引言

§1.1　判别分析研究概况

判别分析是判别样品所属类型的一种多元统计分析方法[1-3]。在生产、科研和日常生活中经常需要根据观测到的数据资料，对所研究的对象进行归属判别。例如在经济学中，根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型；在市场预测中，根据以往调查所得的种种指标，判别下季度产品是畅销、平常或滞销；在地质勘探中，根据岩石标本的多种特性来判别地层的地质年代，由采样分析出的多种成份来判别此地是有矿或无矿，是铜矿或铁矿等；在油田开发中，根据钻井的电测或化验数据，判别是否遇到油层、水层、干层或油水混合层；在农林害虫预报中，根据以往的虫情、多种气象因子来判别一个月后的虫情是大发生、中发生或正常；在体育运动中，判别某游泳运动员的“苗子”是适合练蛙泳、仰泳、还是自由泳等；在医疗诊断中，根据某人多种体检指标（如体温、血压、白血球等）来判别此人是有病还是无病。总之，在实际问题中需要判别的问题几乎到处可见。

§1.1.1　传统的判别分析

传统判别分析内容很丰富，方法很多[1-2]。判别分析按判别的组数来区分，有两组判别分析和多组判别分析；按区分不同总体的所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。按判别分析的准则不同，有马氏距离最小准则、Fisher准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等等。常用的传统判别方法有距离判别法、Fisher判别法、Bayes判别法和逐步判别法。

1 判别分析问题[3]

假定需要作出判别分析的对象分成r类，记作$\mathbf{A}_1$，$\mathbf{A}_2$，$\cdots$，$\mathbf{A}_r$，每一类由m个指

标的n_i个标本确定，即

$$\mathbf{A}_i = \begin{bmatrix} a_{11}^{(i)} & a_{12}^{(i)} & \cdots & a_{1n_i}^{(i)} \\ a_{21}^{(i)} & a_{22}^{(i)} & \cdots & a_{2n_i}^{(i)} \\ \vdots & \vdots & & \vdots \\ a_{m1}^{(i)} & a_{m2}^{(i)} & \cdots & a_{mn_i}^{(i)} \end{bmatrix}_{m\times n_i} ,i = 1,2,\cdots,r \tag{1.1.1}$$

为已知的分类。现在问待判断的对象$x = (x_1, x_2, \cdots, x_m)^T$ 是属于$\mathbf{A}_1$，$\mathbf{A}_2$，…，$\mathbf{A}_r$，中的哪一类？这就构成了判别分析问题的基本内容。

为了能对不同的$\mathbf{A}_1$，$\mathbf{A}_2$，…，$\mathbf{A}_r$做出判别，事先必须有一个一般的规则，一旦知道了x 的值，便能根据这个规则立即做出判断，称这样的一个规则为判别规则。判别规则往往通过某个函数来表达，称为判别函数，记作$W(i;x)$。下面给出有关的符号与说明。

记$n = n_1 + n_2 + \ldots + n_r$，用$a_i, L_i$分别表示第$i$类$\mathbf{A}_i$样本均值和离差矩阵，即

$$a_i = \begin{bmatrix} \bar{a}_1^{(i)} \\ \vdots \\ \bar{a}_m^{(i)} \end{bmatrix}, L_i = \begin{bmatrix} l_{11}^{(i)} & \ldots & l_{1m}^{(i)} \\ \vdots & & \vdots \\ l_{m1}^{(i)} & \ldots & l_{mm}^{(i)} \end{bmatrix}, \bar{a}_j^{(i)} i = 1,2,\cdots,r \tag{1.1.2}$$

其中，$\bar{a}_j^{(i)} = \frac{1}{n}\sum_{k=1}^{n_i} a_{jk}^{(i)}$，$l_{jk}^{(i)} = \sum_{t=1}^{n_i}(a_{jt}^{(i)} - \bar{a}_j^{(i)})(a_{kt}^{(i)} - \bar{a}_{jk}^{(i)})$。并用$x \in \mathbf{A}_i$表示$x$归属于第$i$类$\mathbf{A}_i$。

2 距离判别法[1-10]

距离判别方法就是先建立待判别对象x到第i类$\mathbf{A}_i$的距离$d(x, \mathbf{A}_i)$，然后根据距离最近的原则来判别。即判别函数$W(i;x) = d(x, \mathbf{A}_i)$，判别规则为若$W(k;x) = min\{W(i;x) \mid i = 1,2,\cdots,r\}$，则$x \in \mathbf{A}_i$。

距离$d(x, \mathbf{A}_i)$通常采用印度统计学家拉诺比斯(Mabhalanobis)1936年引入的马氏距离

$$d(x, \mathbf{A}_i) = [(x - a_i)^T V^{-1}(x - a_i)]^{\frac{1}{2}}, V = \frac{L_i}{(n_i - 1)} \tag{1.1.3}$$

3 Fisher判别法[1-10]

Fisher判别方法是基于方差分析的一种判别方法，判别函数$W(x) = u^T x$,其中u为判别系数，计算步骤如下：

(1)计算$L = L_1 + L_2 + \cdots + L_r$，并求出$L^{-1}$。

(2)计算$B=\sum_{i=1}^{r}n_i(a_i-a)(a_i-a)^T$，其中$a=(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m)^T,\bar{a}_j=\frac{1}{n}\sum_{i=1}^{r}n_i\bar{a}_j^{(i)}$。

(3)计算$BL^{(-1)}$的最大特征值对应的特征向量p。特别当$r=2$时，可计算出$p=a_1-a_2$。

(4)计算$L^{(-1)}p,u=L^{(-1)}p$。

为了确定判别规则，先计算$w_i=W(a_i)=u^Ta_i(i=1,2,\cdots,r)$，不妨将$\mathbb{A}_1$,$\mathbb{A}_2$,$\cdots$,$\mathbb{A}_r$重新排序，使得$w_1<w_2<\cdots<w_r$。然后令$c_0=-\infty,c_i=\frac{(w_i+w_{i+1})}{2}$或$c_i=(n_iw_i+n_{i+1}w_{i+1})(n_i+n_{i+1}),c_r=+\infty$。

Fisher准则为，若$c_{k-1}<W(x)<c_k$,则$x\in\mathbb{A}_k$。

4 Bayes判别方法[1-12]

Bayes判别方法的基本思想总是假定对所研究的对象已有一定的认识，常用先验概率来描述这种认识。现在假设r个m维总体密度函数分别为已知的$\phi_i(x)$，且在做判别之前有足够的理由可以认为待判别对象$x\in\mathbf{A}_i$的概率为p_i，如果没有任何这种附加的先验信息，通常取$p_i=\frac{1}{r}$在上述两个假定下，我们将给出一种方便的判别规则，它能使判别概率平均达到最小，这就是Bayes判别方法。

Bayes判别函数$W(i;x)=p_i\phi_i(x)$，判别规则为，若$W(k;x)=max\{W(i;x)\mid i=1,2,\cdots,r\}$,则$x\in\mathbb{A}_i$。

5 逐步判别法[1-10]

逐步判别法与逐步回归法的基本思想类似，都是采用“有进有出”的算法，即逐步引入变量，每引入一个“最重要”的变量进入判别式，同时也考虑较早引入判别式的某些变量，如果其判别能力随新引入变量而变为不显著了（例如其作用被后引入的某几个变量的组合所代替），应及时从判别式中把它剔除去，直到判别式中没有不重要的变量需要剔除，而剩下来的变量也没有重要的变量可引入判别式时，逐步筛选结束。这个筛选过程实质就是作假设检验，通过检验找出显著性变量，剔除不显著变量。由于筛选变量的重要性，近三十年来有大量的文章提出很多种筛选变量的方法。

6 判别效果检验[1-10]

如何知晓判别效果的好坏，这需要对分类的合理性进行假设检验。

选取统计量

$$F=\frac{\sum_{i=1}^{r}n_i(a_i-a)(a_i-a)^T/r-1}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(a_{i_j}-a_i)(a_j^{(i)}-a_i)^T/n-r}\sim F(r-1,n-r)$$

其中，$a=(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m)^T,\bar{a}_j=\frac{1}{n}\sum_{i=1}^{r}n_i\bar{a}_j^{(i)}$，$a_j^(i)=(a_{1j}^{(i)},a_{2j}^{(i)},\cdots,a_{mj}^{(i)})^T$。当$F>F_a(r-1)\times(n-r)$时，说明分类比较合理；否则，说明分类不合理。

§1.1.2 智能化的判别分析

智能化的判别分析是一种机器学习技术，通常是一种搜索。从已知样品通过分类器训练样本，进行判别分析。

有很多学习算法可以用于分类器的训练[13]：

(1)决策树算法[14-16]

它是由J. R. Quinlan于1979年在概念学习系统（Concept Learning System，CLS）算法的基础上提出的。其基本原理是用决策树表示分类的规则。决策树由信息增益（用信息的不确定性的减少作为度量）最大的字段（属性）作为根节点，各个取值为分枝，各个分枝所划分的数据元组为子集，采用递归方法重复建树过程，扩展决策树，最后得到相同类别的子集，再以该类别作为叶节点，从而得到一棵完整的决策树。决策树算法是用于分类和预测的主要技术，决策树学习是以实例为基础的归纳学习算法。它着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则。采用自顶向下的递归方式，在决策树的内部节点进行属性比较，根据不同属性判断从该节点向下的分支，在决策树的叶节点得到结论。所以从根到叶节点就对应着一条合取规则，整棵树就对应着一组析取表达式规则。

(2)候选删除算法[17]

它是利用一般到特殊序的偏序结构，这种结构可以定义在任何概念学习问题中，它提供了一种有用的结构以便于假设空间的搜索。

(3)基于粗集理论的约简算法[18-19]

粗集理论Rough Set的主要思想是在保持分类能力不变的前提下，通过属性约简，导出问题的决策或分类规则。应用粗集理论处理不确定性问题的最显著特点是不需提供问题所需处理的数据集合之外的任何先验信息。属性集的约简(Att ribute Reduct) 是粗集理论中关键的问题之一。所谓约简是属性集的子集，它与原属性集

具有同样的分辨能力。约简反映了一个信息系统的本质信息，求解一个信息系统的全部约简或计算出最佳约简都是NP - 难题。当数据量很大时,应用粗集理论算法十分耗时,因此在有限的时间内求出尽可能短、尽可能好的约简是个启发式搜索问题，也成为一些学者的研究重点。

(4)贝叶斯算法

贝叶斯分类算法是基于贝叶斯定理的分类方法，上一节我们已介绍过。

(5)扩张矩阵算法[20]

扩张矩阵算法是以扩张矩阵理论为基础学习算法，把扩张矩阵中的每一条路径看作一个元素，扩张矩阵可以看作由路径组成的集合，扩张矩阵合并问题是扩张矩阵中路径为元素的集合组的具有公共元素的划分问题，最小扩张矩阵合并问题是最小具有公共路径的扩张矩阵组的划分问题。

(6)k邻近算法[21]

K最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

(7)神经网络[22-23]

人工神经网络是由大量简单的基本元件神经元相互联结，模拟人的大脑神经处理信息的方式，进行信息并行处理和非线形转换的复杂网络系统。人工神经网络处理信息是通过信息样本对神经网络的训练，使其具有人的大脑的记忆、辨识能力，完成各种信息处理功能。该理论具有良好的自学习、自适应、联想记忆、并行处理和非线形转换的能力，避免了复杂数学报导,在样本缺损和参数漂移的情况下，仍能保证稳定的输出。已经成功地应用于众多领域，如模式识别、图像分类、优化计算、人工智能等领域。20世纪90 年代，M. Odom 和R. Sharda 将神经网络模型用于财务预警研究并与传统的多元判别分析进行比较，发现神经网络模型具有更高的判别能力。

(8)遗传算法[24]

遗传算法是模拟自然选择和遗传过程中发生的繁殖、交叉和基因突变现象，在每次迭代中都保留一组候选解，并按某种指标从解群中选取较优的个体，利用遗传

算子(选择、交叉和变异)对这些个体进行组合，产生新一代的候选解群，重复此过程，直到满足某种收敛指标为止。它是有美国Michigan大学J.Holland教授于1975年首先提出来的，并出版了颇有影响的专著《Adaptation in Natural and Artificial Systems》，GA(genetic algorithm)这个名称才逐渐为人所知。随着应用领域的扩展，遗传算法的研究出现了引人注目的新动向：是基于遗传算法的机器学习，这一新的研究课题把遗传算法从历来离散的搜索空间的优化搜索算法扩展到具有独特的规则生成功能的崭新的机器学习算法。这一新的学习机制对于解决人工智能中知识获取和知识优化精炼的瓶颈难题带来了希望。

(9)支持向量机

支持向量机是V.Vapnik等人提出的一种针对分类和回归问题的新型机器学习方法。它基于结构风险最小化原理，能有效地解决过学习问题，具有良好的推广性和较好的分类精确性。传统支持向量机是针对两类分类问题，而在实际应用中，如数据挖掘、文本分类等等，需要处理的数据是海量和多类别的，如何解决大规模多类别的问题，是近几年来研究的重点之一。本文第二章中对其分类理论有详细的阐述。

(10)模糊积分[13,25]

模糊积分是一种基于模糊密度的非线性决策融合方法，它通过定义模糊密度可得到单个分类器或分类器集合的任意子集在融合系统中的重要程度，积分过程转化为分类器所提供的客观结果关于分类器重要程度的积分。它能够很好的体现分类器之间的交互作用，用它作为分类工具已在许多应用领域取得良好的效果。

分类器的设计实际是基于某种知识的分类算法[13]，因此研究分类算法是设计有效分类器的关键。分类问题，它包括模式识别，判别分析。关于分类的研究已有超过一个世纪的历史，已有大量的分类算法问世。但由于分类算法从理论上讲是一个极复杂的函数延拓问题，不存在一种最优的分类算法适用于各种不同的情况，故至今仍有许多分类算法陆续出现，因此也不存在一种最优的判别分析方法适用于各种不同的情况。

§1.2 本文研究的问题及本文的结构

针对支持向量机分类的核函数选择问题，本文提出了新的选择核函数方法—基于粒子群优化算法的组合核函数分类方法，该方法克服了单一核函数不能够精确

描述判别结果的局限性。本文针对某种指标繁多事物的分类判别问题，提出了基于主成分分析和支持向量机的组合判别分析方法，它首先采取多元统计分析中常用的主成分分析方法对这类事物指标进行特征提取，再结合支持向量机分类方法的特点（通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能）对其事物所属种类进行判别分析。通过这种方法，既能简化运算，又可以克服线性方法的局限性，从而提高判断的准确率。本文为方便读者了解该方法，给出了详细的数据分析操作步骤。

本文把基于主成分分析和支持向量机的组合判别分析方法应用到沙尘暴预警问题上，其主要思想是将沙尘暴预警研究视作为一个分类问题，即认为沙尘暴预警的任务就是把沙尘天气预警信号的种类分为三类，分别为黄色，橙色，红色。特别的设计出了判别分析粗框流程图,为大家理解主成分分析和支持向量机组合判别分析方法应用于沙尘暴预警的具体思路。

本文安排如下：

在第二章中，我们首先给出了主成分分析的原理和支持向量机分类的原理，然后提出了基于粒子群优化算法的组合核函数分类方法，最后给出了基于主成分分析和支持向量机的组合判别分析方法及其数据分析步骤。

在第三章中，我们首先谈了关于沙尘暴预警的一些知识，接着作出了一个沙尘暴预警判别分析流程图，最后结合实例数据做了一个判别分析，验证了该方法的有效性。

在第四章中，作了结论报告和一些建议。

第二章　基于主成分分析和支持向量机的组合判别分析

§2.1　主成分分析原理

多指标问题的麻烦一是指标多，二是多指标间的相关性，这样会使它们提供的整体信息发生重叠。这些含有重复信息的大量指标，不但导致问题分析难度增加，而且会歪曲事物的真正特征及其发展规律。主成分分析(Principal Component Analysis) 也称主分量分析,是由霍特林(Hotelling) 于1933 年首先提出的。它是一种将原来多个指标转化为少数几个互不相关综合指标的数据降维法。综合指标是原来多个指标的线性组合，虽然这些综合指标是不能直接观测到的，但这些综合指标间互不相关，又能反映原来多指标的信息，达到了减少指标和删除重复信息的目的[1-2]。

§2.1.1　主成分分析的数学原理

主成分分析的数学原理是基于代数学的矩阵理论，其理论如下叙述[1-10,26]。

对于有p个指标的总体$X = (X_1, X_2, \ldots, X_p)'$主成分分析方法确立$X$的综合指标$Y_1, Y_2, \ldots, Y_m, m \leq p$的思想如下：

(1)Y_i是X的线性组合，即要求$Y_i = l'X$，l_i是$p \times 1$维待定的单位向量，$i = 1, 2, \ldots, m$。

(2)Y_1是$X_1, X_2, \ldots, X_p$的一切线性组合方差最大的；Y_2是与Y_1不相关的且是$X_1, X_2, \ldots, X_p$的一切线性组合中方差最大的；依此类推，Y_m与$Y_1, Y_2, \ldots, Y_{m-1}$都不相关，是$X_1, X_2, \ldots, X_p$的一切线性组合中方差最大的。这样的$Y_1, Y_2, \ldots, Y_m$称为$X$的第一，第二，...，第$m$主成分。方差大小表示包含原有信息的多少，因此，$Y_1, Y_2, \ldots, Y_m$包含的信息量依次递减。在实际应用中，选取前几个主成分，虽然损失了一定的信息，但抓住了主要矛盾，简化了分析。

求解主成分的定理如下：

定理1：设A为P阶对称矩阵，其特征根为$\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_p$各个特征根对应的单位化特征向量为$\gamma_1, \gamma_2, \ldots, \gamma_p$，对一个待定的$P$维单位向量$l$，则有

①当$l = \gamma_1$时，$\max(l'Al)$的最大值为λ_1。

②当$l_{k+1} = \gamma_{k+1}$且$l'_{k+1}l_i = 0$时，$i = 1, 2, \ldots, k$，$\max(l'_{k+1}Al_{k+1})$的最大值

为λ_{k+1}。

定理2：对于有p个指标的总体$X=(X_1,X_2,\ldots,X_p)'$其协方差矩阵$\sum>0$，其特征根为$\lambda_1,\lambda_2,\ldots,\lambda_p$，其中$\lambda_1\geq\lambda_2\geq\cdots,\geq\lambda_m>0$，$\lambda_{m+1}=\lambda_{m+2}=,\ldots,=\lambda_p=0$；$\gamma_i$为对应的单位化特征向量，则第$i$个主成分为$Y_i=\gamma_i'X,i=1,2,\ldots,m$。

根据上述主成分求解定理，这样寻找总体X的主成分就转化为求X的协方差矩阵$\sum$的特征根和特征向量的问题。求解X的协方差矩阵$\sum$的特征根和特征向量步骤如下：

① 求X的协方差矩阵$\sum$的特征根，记为$\lambda_1\geq\lambda_2\geq\cdots,\geq\lambda_m,\lambda_{m+1}=\lambda_{m+2}=,\ldots,=\lambda_p=0$。

② 求λ_i对应单位化特征向量$\gamma_i,i=1,2,\ldots,m$。

③ 计算第i个主成分$Y_i=\gamma_i'X,i=1,2,\ldots,m$。

§2.1.2 主成分的精度分析

在实际应用中，常常选取前几个主成分，使其总体累计方差达到一定的精度[26]，这就是下面要说明的主成分贡献率。

设总体$X=(X_1,X_2,\ldots,X_p)'$的协方差矩阵为$\sum=(v_{ij})_{p\times p}$，其秩$R(\sum)=m$。$\lambda_1,\lambda_2,\ldots,\lambda_p$是$\sum$的$p$个特征根，且$\lambda_1\geq\lambda_2\geq\cdots,\geq\lambda_m,\lambda_{m+1}=\lambda_{m+2}=,\ldots,=\lambda_p=0$，特征向量$\gamma_1,\gamma_2,\ldots,\gamma_p$是$\lambda_1,\lambda_2,\ldots,\lambda_p$对应的标准正交向量。$X$的第$i$个主成分$Y_i=\gamma_i'X$，$i=1,2,\ldots,m$。则称：

① $\lambda_i/\sum_{j=1}^{p}\lambda_j$为主成分$Y_i$的贡献率。

② $\sum_{k=1}^{i}\lambda_i/\sum_{j=1}^{p}\lambda_j$为主成分$Y_1,Y_2\cdots,Y_i$的累计贡献率。

在实际应用中，选取前几个主成分，使其累计贡献率达到70%～90%即可。

§2.2 支持向量机分类

传统的基于统计学的方法多数建立在大数定理这一理论基础上的渐进理论，要求学习样本数目足够多[27]。但在实际应用中，这一前提往往不切实际，所以在小样本情况下，很难取得理想的学习效果和泛化性能。V.Vapnik教授等人针对小样本情况下的机器学习，建立了统计学习理论，并在此基础上提出了支持向量机方

法。V.Vapnik等人从六、七十年代开始致力于此方面研究，到九十年代中期，随着其理论的不断发展和成熟，也由于神经网络等学习方法在理论上缺乏实质性进展，统计学习理论开始受到越来越广泛的重视。近年来在其理论研究和算法实现方面都取得了飞速发展，开始成为克服维数灾难和过学习等传统困难的有力手段。统计学习理论（Statistical Learning Theory或SLT）是一种专门研究小样本情况下机器学习规律的理论，该理论针对小样本统计问题建立了一套新的理论体系，在这种体系下的统计推理规则不仅考虑了对渐近性能的要求，而且追求在现有有限信息的条件下得到最优结果[30-31]。

统计学习理论的主要研究内容包括[27,32]：

1. VC维

为了研究经验风险最小化函数集的学习一致收敛速度和推广性，统计学习理论定义了一些指标来衡量函数集的性能，其中最重要的就是VC维(Vapnik-Chervonenkis Dimension)。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，则称函数集能够把h个样本打散。函数集的VC维就是能够打散的最大样本数目。如果对任意的样本数，总有函数能打散它们，则函数集的VC维就是无穷大。一般而言，VC维越大，学习机器的学习能力越强，但学习机器也越复杂。目前还没有通用的关于计算任意函数集的VC维的理论，只有对一些特殊函数集的VC维可以准确知道。例如，N维实数空间中线性分类器和线性实函数的VC维是$n+1$，$sin(ax)$的VC维为无穷大。对于给定的学习函数集，如何用理论或实验的方法计算其VC维是当前统计学习理论研究中有待解决的一个重要问题。

2. 推广性的界

统计学习理论系统地研究了经验风险和实际风险之间的关系，也即推广性的界。根据统计学习理论中关于函数集推广性的界的理论，对于指示函数集中所有的函数，经验风险$R_{emp}(\omega)$和实际风险$R(\omega)$之间至少以概率$1-\eta$满足如下关系：

$$R(\omega) \leq R_{emp}(\omega) + \sqrt{\frac{h(ln\frac{2n}{h}+1) - ln(\frac{\eta}{4})}{n}}$$

其中，h是函数集的VC维，n为样本数。

由上述经验风险$R_{emp}(\omega)$和实际风险$R(\omega)$之间的关系公式可知，学习机器的实际风险由两部分组成：(1)训练样本的经验风险；(2)置信范围(同置信水平$1-\eta$有关，

且与学习机器的VC维和训练样本数有关)。即

$$R(\omega) \leq R_{emp}(\omega) + \Phi(\frac{h}{n})$$

它表明，在训练样本有限的情况下，学习机器的VC维越高，则置信范围就越大，导致实际风险与经验风险之间可能的差就越大。在设计分类器时，不但要使经验风险最小化，还要使VC维尽量小，从而缩小置信范围，使期望风险最小化。寻找反映学习机器能力的更好参数从而得到更好的界是今后学习理论的重要研究方向之一。

3 结构风险最小化原则

传统机器学习方法中普遍采用的经验风险最小化原则在样本数目有限时是不合理的，因此，需要同时最小化经验风险和置信范围。统计学习理论提出了一种新的策略，即把函数集构造为一个函数子集序列，使各个子集按照VC维的大小排列；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，取得实际风险的最小化，这种思想被称作结构风险最小化(Structural Risk Minimization)，实现SRM原则可以有两种思路：(1)在每个子集中求最小经验风险，然后选择使最小经验风险和置信范围之和最小的子集。(2)设计函数集的某种结构使每个子集中都能取得最小的经验风险，然后，只需选择适当的子集使置信范围最小，则这个子集中使经验风险最小的函数就是最优函数。支持向量机方法实际上就是第二种思路的实现。

支持向量机(Support Vector Machines,SVM)是针对分类和回归问题的统计学习理论，主要思想可概括为两点[27]：

(1) 针对线性可分情况进行分析。对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。

(2) 基于结构风险最小化理论之上，在特征空间中建构最优分割超平面，使得学习机得到全局最优化，并在整个样本空间的期望风险以某个概率满足一定的上界。

§2.2.1 支持向量机分类理论

支持向量分类理论已被成功的应用到图像分类和医疗领域的判别分析等多个领域。它的理论如下[27-44]：

对训练集(x_1, y_1)，(x_2, y_2)，$\cdots$，(x_n, y_n)，$x \in \mathbb{R}^d$，$y \in \{-1, 1\}$。如果x_i属于第

一类，则标记为1，如果属于第二类，则标记为-1，学习的目的是要构造一个判别函数$f(x)$，将两类尽可能的分开。

如果n个训练样本是线性可分的，则必然存在某个超平面

$$(w \cdot x) + b = 0 \tag{2.2.1}$$

将两类样本完全分开。构造并求解对变量$w \in \mathbb{R}^d$以及$b \in \mathbb{R}$的最优化问题：

$$\begin{cases} minimize: \frac{1}{2}\|w\|^2 \\ subject\ to: y_i[(w \cdot x_i) + b] \geq 1, i = 1, 2, \cdots, n \end{cases} \tag{2.2.2}$$

判别函数为：

$$f(x) = sgn[(w \cdot x) + b] \tag{2.2.3}$$

为了求解式(2.2.2)，利用Lagrange 优化方法可以把上述最优超平面问题转化为其对偶问题：

$$\begin{cases} maximize: \sum\limits_{i=1}^{n} a_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) \\ subject\ to: \sum\limits_{i=1}^{n} a_i y_i = 0, a_i \geq 0, i = 1, 2, \cdots, n \end{cases} \tag{2.2.4}$$

其中a_i为Lagrange乘子。

式(2.2.4) 是一个不等式约束下二次函数寻优的问题，存在唯一解。解中只有一部分(通常是少部分)a_i不为0，对应的样本就是支持向量(Support Vector，SV)。这样判别函数成为：

$$f(x) = sgn[\sum_{i \in SV} a_i y_i (x_i \cdot x) + b] \tag{2.2.5}$$

上述问题都是局限于完全线性可分的情况下研究的，如果线性不可分，可引入松弛变量ξ和惩罚因子C，把上述理论推广到广义的最优超平面求解如下：

$$\begin{cases} minimize: \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{n} \xi_i \\ subject\ to: y_i[(w \cdot x_i) + b] \geq 1 - \xi_i, i = 1, 2, \cdots, n \end{cases} \tag{2.2.6}$$

式中ξ_i是对错误分类误差的度量，C是预先设定的数，其中$C \geq 0$，控制对错分样本惩罚的程度。式(2.2.6)的对偶问题：

$$\begin{cases} maximize: \sum\limits_{i=1}^{n} a_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) \\ subject\ to: 0 \leq a_i \leq C, i = 1, 2, \cdots, n, \sum\limits_{i=1}^{n} a_i y_i = 0 \end{cases} \tag{2.2.7}$$

在数据高度线性不可分的情况下，即使引入松弛变量和惩罚因子仍然不能实现分类。针对这种情况一种思路就是把数据映射到高维特征空间，使得在高维空间里，这些数据线性可分。选取适当的映射$\Phi: x \longrightarrow \Phi(x)$，通过$\Phi(x)$把数据映射到高维特征空间使其线性可分，就可以利用已经得到的线性可分的学习方法，求解下面的优化问题。

$$\begin{cases} maximize: \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \Phi(x_i)\cdot\Phi(x_j) \\ subject\ to: 0 \le a_i \le C, i = 1, 2, \cdots, n, \sum_{i=1}^{n} a_i y_i = 0 \end{cases} \tag{2.2.8}$$

按照这个思路，核心问题是如何选择合适的$\Phi(x)$实现从数据空间到高维的特征空间的映射。这个问题很复杂，目前还没有一套完整的理论来来指导映射函数的选择。但是如果注意到优化问题中仅仅出现了$(\Phi(x_i)\cdot\Phi(x_j))$，即$\Phi(x)$的内积，问题就简单了。如果能够找到一个函数$K$，满足

$$K(x_i, x_j) = (\Phi(x_i)\cdot\Phi(x_j)) \tag{2.2.9}$$

则$K(\cdot)$就隐式定义了一个从数据空间到高维特征空间的映射，而K就被称为核函数。所以如果这n个训练样本是线性不可分的，可以将原问题通过核函数映射到高维空间，这样最优化问题就变为：

$$\begin{cases} maximize: \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \\ subject\ to: 0 \le a_i \le C, i = 1, 2, \cdots, n, \sum_{i=1}^{n} a_i y_i = 0 \end{cases} \tag{2.2.10}$$

这里$K(x_i, x_j) = (\Phi(x_i)\cdot\Phi(x_j))$。判别函数就变成：

$$f(x) = sgn[\sum_{i\in SV} a_i y_i K(x_i \cdot x) + b] \tag{2.2.11}$$

式2.2.10主要是选取适当的核函数K和适当的参数C，这种支持向量分类机被称为$C-$支持向量分类机$(C-SVC)$。$C-$支持向量分类机有两个矛盾的目标：最大化间隔和最小化训练错误。其中的参数C起着调和这两个目标的作用。定性的讲，C值有着明确的含义：选取最大的C值，意味着更强调最小化训练。但定量的讲C值本身没有实际意义，这意味着C值选取较困难。对此，人们提出了一个改进方法：$v-$支持向量分类机$(v-SVC)$。

$v-SVC$的原始问题：

$$\begin{cases} minimize: \frac{1}{2}\|w\|^2 - v\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ subject\ to: y_i[(w\cdot x_i)+b] \geq \rho - \xi_i, \\ \qquad\qquad \rho \geq 0, \xi_i \geq 0, i = 1,2,\cdots,n \end{cases} \tag{2.2.12}$$

这里不含参数C，而是换成了参数v另外还有一个参数ρ。其对偶问题变为：

$$\begin{cases} maximize: -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j (x_i, x_j) \\ subject\ to: i = 1,2,\cdots,n, \sum_{i=1}^{n} a_i y_i = 0 \\ \qquad\qquad 0 \leq a_i \leq \frac{1}{n}, i = 1,2,\cdots,n \\ \qquad\qquad \sum_{i=1}^{n} a_i \geq v \end{cases} \tag{2.2.13}$$

引进核函数

$$K(x_i, x_j) = (\Phi(x_i)\cdot\Phi(x_j)) \tag{2.2.14}$$

式(2.2.12)的对偶问题可表示为

$$\begin{cases} maximize: -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \\ subject\ to: i = 1,2,\cdots,n, \sum_{i=1}^{n} a_i y_i = 0 \\ \qquad\qquad 0 \leq a_i \leq \frac{1}{n}, i = 1,2,\cdots,n \\ \qquad\qquad \sum_{i=1}^{n} a_i \geq v \end{cases} \tag{2.2.15}$$

其判别函数仍为：

$$f(x) = sgn[\sum_{i\in SV} a_i y_i K(x_i\cdot x) + b] \tag{2.2.16}$$

SVM中不同的内积核函数将形成不同的算法。在实际应用中，广泛的使用的有四类：线性核函数，多项式核函数，径向基核函数和$Sigmoid$函数。

① 线性核函数

$$K(x_i, x_j) = (x_i\cdot x_j) \tag{2.2.17}$$

② 多项式函数

$$K(x_i, x_j) = (x_i\cdot x_j + 1)^d \tag{2.2.18}$$

③径向基核函数

$$K(x_i,x_j)=e^{-\|x_i-x_j\|/2\sigma^2} \tag{2.2.19}$$

④ *Sigmoid*核函数

$$K(x_i,x_j)=tanh(W(x_i\cdot x_j)+b) \tag{2.2.20}$$

§2.2.2 基于粒子群优化算法的组合核函数分类方法

根据上述理论可知，在数据高度线性不可分，即使引入松弛变量和惩罚因子仍然不能实现分类的情况下，采取选择核函数的办法来实现分类。那么怎样选择核函数才能使结果最优呢？这是我们这一小节要解决的问题。本节采取粒子群优化组合核函数的办法来选择核函数使结果达到最优。

1 组合核函数模型[45,46]

组合核函数模型是基于一个基本假设，即只用一种核函数模型不能够精确地描述出判别过程和判别结果，而各种模型综合起来，就可以从不同的侧面反映整个判别过程，从而使判别结果更加精确，以实现核函数之间的取长补短。

设用m种不同的核函数进行测试，则由这m 个单一核函数构成的组合模型为:

$$\widetilde{K}(x_i,x_j)=\sum_{d=1}^{m}w_dK_d(x_i,x_j) \tag{2.2.21}$$

其中,$\widetilde{K}(x_i,x_j)$为组合的核函数，组合的核函数仍是核函数[46]，$K_d(x_i,x_j)$为第d个单一核函数。w_d 为第d 个核函数的权重，且满足:

$$\sum_{d=1}^{m}w_d=1,w_d\geq 0,d=1,2,\cdots,m \tag{2.2.22}$$

2 组合核函数分类方法

式(2.2.10)最优化问题就改进为：

$$\begin{cases} maximize:\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_j\widetilde{K}(x_i,x_j) \\ subject\ to:0\leq a_i\leq C,i=1,2,\cdots,n,\sum_{i=1}^{n}a_iy_i=0 \end{cases} \tag{2.2.23}$$

这里$\widetilde{K}(x_i,x_j)=\sum_{d=1}^{m}w_dK_d(x_i,x_j)$。其判别函数为：

$$f(x)=sgn[\sum_{i\in SV}a_iy_i\widetilde{K}(x_i\cdot x)+b] \tag{2.2.24}$$

求解上述组合的核函数的关键是确定各单一核函数的权重，使组合核函数模型能更有效地提高分类精度。

传统的确定加权系数的方法有算术平均方法、预测误差平方和倒数方法、均方误差导数方法、简单加权平均法、二项式系数法等[45]。但都有各自的不足，例如，算术平均方法对每个模型平等对待,没有考虑模型重要程度的差异；误差法中的权重则完全依赖于各个模型所得的误差,实际应用中很难达到理想的要求；近年来,一些非参数方法和智能算法兴起,这些方法具有较好的鲁棒性，所以在组合模型中的应用越来越受到人们的重视。本文通过智能算法—微粒群算法对单一核函数的权重进行求解。

3 粒子群算法[48-50]

粒子群优化算法(Particle Swarm Optimization , PSO)是由Kennedy和Eberhart受鸟群觅食行为的启发，于1995年提出的。它同遗传算法类似，是一种基于迭代的优化工具。但它没有遗传算法的交叉和变异操作，而是类似梯度下降算法使各染色体向适应值最高的方向群游。与其它的优化算法相比，粒子群算法不仅具有全局寻优能力，而且参数少，容易实现。粒子群算法的基本思想是将优化问题的每一个解称为粒子，定义一个适应值函数来衡量每个粒子的优越程度。每个粒子根据自己和其它粒子的“飞行经验”，达到从全空间搜索最优解的目的。

PSO 初始化为一群随机粒子(随机解)。然后通过叠代找到最优解。在每一次叠代中，粒子通过跟踪两个“极值”来更新自己。第一个就是粒子本身所找到的最优解。这个解叫做个体极值pbest。另一个极值是整个种群目前找到的最优解。这个极值是全局极值gbest。另外也可以不用整个种群而只是用其中一部分最为粒子的邻居，那么在所有邻居中的极值就是局部极值。

在找到这两个最优值时，粒子根据如下的公式来更新自己的速度和新的位置：

$$v_{ij}^{k+1} = v_{ij}^{k} + c_1 * r_1 * (pbest_{ij}^{k} - x_{ij}^{k}) + c_2 * r_2 * (gbest_{j}^{k} - x_{ij}^{k}) \tag{2.2.25}$$

$$x_{ij}^{k+1} = x_{ij}^{k} + v_{ij}^{k+1} \tag{2.2.26}$$

其中，i代表第i个粒子，j代表速度或位置的第j维，k代表迭代代数，如v_{ij}^{k}和x_{ij}^{k}分别是第i粒子P_i在第k次迭代中第j维的速度和位置，两者均被限制在一定的范围内；c_1和c_2是学习因子，通常$c_1, c_2 \in [0,4]$；r_1和r_2是介于$[0,1]$之间的随机数；$pbest_{ij}^{k}$是粒子P_i在第j维的个体极值的坐标；$gbest_{j}^{k}$是群体在第j维的全局极值的坐标。式(2.2.25)和

式(2.2.26)组成基本PSO算法公式。

4 粒子群优化算法用于求解组合核函数加权系数w_d

确定加权系数w_d的PSO 算法步骤如下：

(1) 初始化粒子群，随机产生N个符合条件的粒子，该粒子是由核函数系数组成的d维向量。

(2) 对样本进行测试，计算每个个体的适应值，以反映本组合核函数模型的推广分类能力,适应值函数如下：

$$\begin{cases} f = maximize(\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \sum_{d=1}^{m} w_d K_d(x_i, x_j)) \\ subject\ to: 0 \le a_i \le C, i = 1, 2, \cdots, n, \sum_{i=1}^{n} a_i y_i = 0 \\ \sum_{d=1}^{m} w_d = 1, w_d \ge 0, d = 1, 2, \cdots, m \end{cases} \tag{2.2.27}$$

(3) 按照(2.2.27) 式计算的适应函数值f 与自身的最优值$f(pbest)$进行比较:如果$f < f(pbest)$，则用新的适应值取代前一轮的优化解，用新的粒子取代前一轮的粒子，即：

$$f(pbest) = f$$

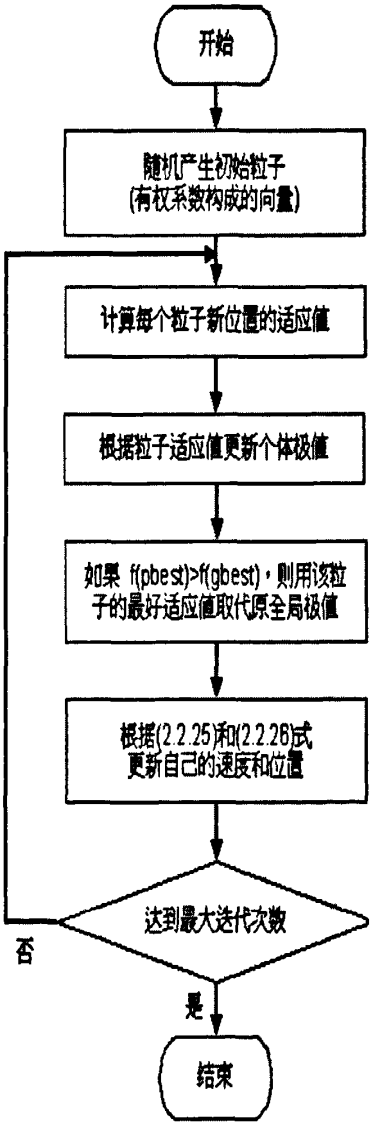
(4) 将每个粒子的最好的适应值$f(pbest)$ 与所有粒子的最好适应值$f(gbest)$ 进行比较：如果$f(pbest) > f(gbest)$，则用该粒子的最好适应值取代原有全局最好适应值,同时保存粒子的当前状态。

(5) 判断适应值是否满足要求，如不满足要求，再进行新一轮的计算，按(2.2.25)和(2.2.26)式将粒子进行移动，从而产生新的粒子，返回步骤(2)；如果适应值满足要求，计算结束。

用上述PSO 算法求出w_d，利用(2.2.23)、(2.2.24) 式即可求出最终的分类结果。由于本人水平有限，该算法还未在计算机上实现。

求解组合核函数加权系数w_d问题的粒子群优化算法的流程如图2.1所示：

图 2.1: 粒子群优化权系数w_d流程图



§2.2.3 多分类支持向量机

上述理论是针对两分类问题提出的，而实际应用中需要解决的是多分类问题。因此将两分类SVM推广到多分类问题很有必要。下面以k分类问题为例，给出多类SVM分类模型[27-28]。k分类问题可表述为：根据给定的N个独立同分布样本$\{(x_i, y_i), i = 1, 2, \cdots, N\}$，$x_i \in \mathbb{R}^d$，$y_i \in \{i = 1, 2, \cdots, k\}$ 表示x_i的类别，确定最优函数$f(x, \omega_0)$对该依赖关系进行估计，使风险最小。大体上有两种途径将SVM推广到多分类问题，直接法和分解法。

(1) 直接法: 将多个SVM分类面的参数求解合并到一个最优化问题中，通过求解最优化问题直接实现多分类。

(2) 分解法: 通过某种方式构造一系列标准的两类SVM分类器，并将它们组合在一起来实现多分类。即分解法将多类别分类分成一系列两类问题的子问题。

§2.2.4 模型选择问题

在实际应用中，要把它转化为能用支持向量机求解的数学模型。这项工作称为模型选择，它包括[28]：

① 选取训练集$T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$；

② 选择某一支持向量机，如$C-$支持向量分类机，或$v-$支持向量分类机。

③ 选取支持向量机中的核函数和其中的参数，如C，v。

§2.3 主成分分析和支持向量机的组合判别分析

基于主成分分析和支持向量机的组合判别分析方法思想是：采取多元统计分析中常用的主成分分析的方法对事物指标进行特征提取，对被提取特征后的指标数据作为样本数据进行SVM分类，来对其事物所属种类进行判别分析。通过这种方法，既能简化运算，又可以把原先的线性方法变为非线性方法，从而提高判断的准确率。

§2.4 数据分析步骤

对数据进行主成分分析我们采用SPSS软件工具求解主成分，SPSS是世界公认的标准统计软件之一，集数据录入、资料编辑、数据管理、统计分析、报表制作、图

形绘制为一体。对于支持向量机，它的软件工具主要有LIBSVM和SVMLight，我们采用LIBSVM[51]。LIBSVM 是一个开源的软件包，是台湾大学林智仁博士等开发的，可以解决上面所提到的三类机器学习基本问题，解决分类问题（包括C- SVC、v-SVC）、回归问题（包括e-SVR、v-SVR ）以及分布估计（one-class-SVM ）等问题，提供了线性、多项式、径向基和Sigmoid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。另外还提供了WINDOWS 平台下的可视化操作工具SVM-toy，并且在进行模型参数选择时可以绘制出交叉验证精度的等高线图。

特别的强调一下，本文主要是应用到它的分类功能，来进行判别分析。下面我们给出主成分分析和支持向量机组合判别分析方法具体应用到实际情况中的数据分析步骤：

①首先将数据录入SPSS数据文件，接着打开其数据文件，依次点选Analysis $\longrightarrow$ Data Reduction $\longrightarrow$ Factor，进入Factor Analysis (因子分析) 对话框。

②使用SPSS统计软件做主成分分析。选取前几个主成分，使其累计贡献率达到70% $\sim$ 90%即可。

③把选中的前几个主成分按照LIBSVM软件包所要求的格式准备数据集。LIBSVM使用的训练数据和测试数据文件格式如下：

$$< label >< index1 >:< value1 >< index2 >:< value2 > \cdots$$

其中label 是训练数据集的目标值，对于分类，它是标识某类的整数（支持多个类）；对于回归，是任意实数。index 是以1 开始的整数，表示特征的序号；value为实数，也就是我们常说的特征值或自变量。当特征值为0 时，特征序号与特征值value都可以同时省略，即index可以是不连续的自然数。label 与第一个特征序号、前一个特征值与后一个特征序号之间用空格隔开。测试数据文件中的label 只用于计算准确度或误差，如果它是未知的，只需用任意一个数填写这一栏，也可以空着不填。例如：

1 1 : 0.2 2 : 1.3 3 : 0.5

2 1 : 0.3 2 : 1.2 3 : 0.6

3 1 : 0.4 2 : 1.1 3 : 0.4

④对数据进行简单的缩放操,svmscale命令是把每个特征的值映射到[0,1]或者[-1,1]上面。首先应该明白，libsvm里面的svmtrain命令是不要求输入的样本矩阵的每

个值一定要在[0,1]或者[-1,1]范围的。对数据集进行缩放的目的在于：1）避免一些特征值范围过大而另一些特征值范围过小；2）避免在训练时为了计算核函数而计算内积的时候引起数值计算的困难。因此，通常将数据缩放到[0,1]或者[-1,1]之间。用法：

$$svmscale\ filename$$

⑤考虑选用SVM类型（C- SVC,v-SVC）和核函数，对整个训练集进行训练获取支持向量机模型。svmtrain命令实现对训练数据集的训练，获得SVM模型。用法：

$$svmtrain\ [options]\ training\ set\ [model\ file]$$

其中，options（操作参数）：可用的选项即表示的涵义如下所示。

-s SVM类型：设置SVM 类型，默认值为0，可选类型有：

$$0 \longrightarrow C-SVC$$

$$1 \longrightarrow v-SVC$$

-t 核函数类型：设置核函数类型，默认值为2，可选类型有：

0 ⟶线性核函数

1 ⟶ 多项式函数

2 ⟶ 径向基核函数

3 ⟶ *Sigmoid*核函数

⑥利用获取的模型进行测试与预测判别。在第六步对整个训练集进行训练获取支持向量机模型的基础上，采用svmpredict 命令可以得出待测数据的判别分析结果。用法：

$$svmpredict\ [options]\ test\ file\ model\ file\ \ output\ file$$

-b 概率估计预测：设置是否需要进行概率估计预测，可选值为0或者1，默认值为0。model file是由svmtrain产生的模型文件；test file 是要进行预测的数据文件；output file是svmpredict的输出文件，表示预测的结果值。

第三章　在沙尘暴预警中的应用

§3.1　沙尘暴预警

沙尘暴是一种极具破坏力的自然灾害,它可造成房屋倒塌、交通供电受阻或中断、火灾、人蓄伤亡等,污染自然环境,破坏作物生长,给国民经济建设和人民生命财产安全造成严重的损失和极大的危害。

表 3.1: 沙尘暴预警信号及防御指南

黄	12小时内可能出现沙尘暴天气（能见度小于1000米）；或者已经出现沙尘暴天气，并可能持续。	1.有关部门根据情况启动防御工作预案; 2.做好防风防沙准备，及时关闭门窗； 3.注意携带口罩、纱巾等防尘用品，以免沙尘对眼睛和呼吸道造成损伤； 4. 做好精密仪器的密封工作； 5.固紧门窗、围板、棚架、户外广告牌、临时搭建物等易被风吹动的搭建物,妥善安置易受沙尘暴影响的室外物品。
橙	6小时内可能出现强沙尘暴天气（能见度小于500米）；或者已经出现强沙尘暴天气，并可能持续。	1.有关部门根据情况启动防御工作预案； 2.用纱巾蒙住头防御风沙的行人要保证有良好的视线，注意交通安全； 3.注意尽量少骑自行车，刮风时不要在广告牌、临时搭建物和老树下逗留，驾驶人员注意沙尘暴变化，小心驾驶； 4.机场、高速公路、轮渡码头等应注意交通安全； 5.各类机动交通工具应采取有效措施保障安全。 其他同沙尘暴黄色预警信号。
红	6小时内可能出现强沙尘暴天气（能见度小于50米）；或者已经出现强沙尘暴天气，并可能持续。	1.有关部门根据情况启动防御工作预案； 2.应急处置与抢险单位随时准备启动抢险应急方案；人员应呆在防风安全的地方，不要在户外活动； 3.受特强沙尘暴影响地区的机场暂停飞机起降，高速公路和轮渡码头等暂时封闭或者停航。 其他同沙尘暴橙色预警信号。

图 3.1: 判别分析流程粗框图

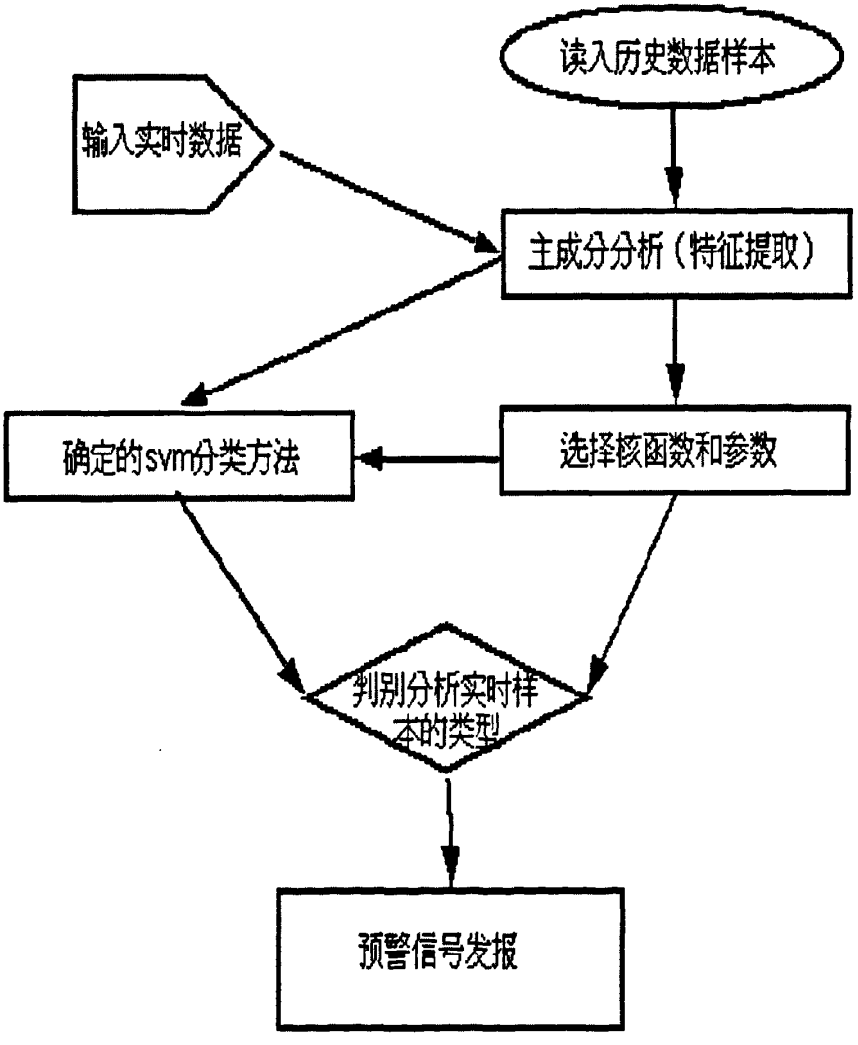


表3.1是来自国家气象局对沙尘暴预警信号的制定标准及防御指南，我们下面的沙尘暴种类的划分依据此表。图3.1是为大家理解主成分分析和支持向量机组合判别分析方法应用于沙尘暴预警的具体思路，给出的粗框流程图,我们从图中不难理解和分析出该方法应用在此处的步骤。

近十几年来，沙尘暴频频发生，沙尘暴这个名词不时见诸于新闻媒体，越来越受到相关国家人民和领导人的重视。科学发展观也教育和指导我们要有科学忧患意识，很有必要加强沙尘暴预警的研究。

近年来对沙尘天气的研究已有很多，其中主要是集中在沙尘天气的卫星遥感监测、沙尘天气过程的天气学分析、沙尘天气的气候学特征分析等[52]，另外也有许多研究是通过数值预报模式对沙尘天气进行个例分析。本文研究工作是已有的研究方法和成果的基础上，利用气象数值预报给出的气象指标，结合当地沙尘暴历史数据指标进行判别分析是哪种程度的沙尘暴，也即是说判别分析当前天气指标做出红或黄或橙的沙尘暴预警信号，做好预警工作，有效地提醒相关人员做好防御准备。

§3.2 实例分析

甘肃河西走廊沙区是我国沙尘暴的多发区，民勤县又是河西走廊沙尘暴的重灾区。民勤县地处巴丹吉林和腾格里两大沙漠的接壤地带，位于河西走廊东北部荒漠区，当地气候干旱少雨，沙尘暴频繁[53]。

表3.2是来自1998年至2001年民勤地区发生34次沙尘暴的前期气象指标数据，其中，用x_1表示发生日期，用x_2表示前7天的平均气温(c)，用x_3表示日均升高的温度(c)，用x_4表示表示前7天的平均气压(hPa)，用x_5表示日均下降气压(hPa)，用x_6表示前期风速(m/s)，用x_7表示前期降水(mm/d)，用x_8表示最大风速(m/s)，用x_9表示持续小时$(h:m)$，用x_{10}表示能见度(m)。

表 3.2: 民勤沙尘暴资料

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
02/16/1998	1.44	2.15	860.63	6.70	5.10	0.009	14.8	0:22	500
04/15/1998	11.99	1.40	862.56	0.73	3.62	0.810	25.2	6:38	100
04/19/1998	15.64	1.76	862.26	1.72	2.65	0.579	11.5	1:31	200
04/27/1998	15.99	3.48	857.00	3.95	3.75	0.368	14.6	2:15	600
04/30/1998	17.58	0.55	858.54	1.30	3.78	2.300	20.6	1:22	400
05/20/1998	19.20	2.23	858.54	1.94	4.36	0.960	17.9	0:29	800
11/01/1998	5.96	1.90	868.91	3.30	6.03	0.220	23.1	3:40	600
02/17/1999	-0.34	4.20	870.24	3.50	4.37	0.034	19.0	0:56	900
03/15/1999	4.59	3.00	859.19	5.25	3.77	0.029	18.3	1:03	300
04/04/1999	7.04	2.70	864.64	0.97	5.43	0.025	20.4	3:18	600
04/10/1999	13.43	2.07	860.04	2.18	5.02	0.025	22.1	3:59	200
07/10/1999	20.76	1.63	857.84	1.83	2.40	4.925	21.0	0:38	700
08/05/1999	28.88	1.10	854.20	2.13	5.92	0.635	13.8	1:15	700
11/23/1999	1.71	2.30	865.43	3.30	5.26	0.030	22.5	1:38	700
03/21/2000	5.56	1.00	860.13	5.50	2.13	0.058	16.8	0:07	900
03/26/2000	4.83	1.53	868.16	1.83	3.27	0.048	20.0	1:55	300
04/09/2000	8.71	2.13	865.27	2.03	6.04	0.033	18.0	0:26	700
04/12/2000	10.06	4.50	863.51	7.15	2.12	0.031	24.0	2:23	50
06/05/2000	21.29	1.30	857.47	1.15	2.49	0.067	20.0	4:51	30
07/06/2000	23.51	2.05	857.84	0.20	3.36	0.375	18.0	2:59	50
08/01/2000	23.02	4.00	859.12	4.40	2.48	0.450	18.0	1:58	700
10/17/2000	6.86	2.50	870.29	4.40	2.10	0.730	16.4	1:10	900
11/18/2000	-1.13	2.70	869.24	5.30	4.48	0.400	20.0	3:08	400
12/23/2000	-2.17	0.85	868.31	3.07	3.55	0.041	20.0	0:40	100
12/31/2000	-3.06	1.40	867.97	3.30	2.28	0.034	21.0	4:04	100
01/02/2001	-3.57	0.50	867.26	2.35	3.79	0.033	18.5	0:11	800
01/12/2001	-5.03	2.55	863.30	3.53	2.59	0.028	19.0	1:50	600
03/04/2001	0.412	0.10	868.84	8.50	3.25	0.015	18.0	1:24	700
03/19/2001	3.43	2.20	865.41	0.76	6.02	0.013	21.0	3:44	300
04/06/2001	10.96	0.90	862.40	5.80	2.30	0.011	18.0	2:01	300
04/08/2001	10.16	1.00	861.34	3.73	1.98	0.011	22.0	6:06	100
04/09/2001	8.11	1.00	863.54	3.73	7.80	0.011	18.5	1:46	700
04/12/2001	5.50	3.90	866.26	8.65	4.53	0.010	19.0	1:15	600
06/11/2001	22.63	1.70	856.60	2.10	2.67	0.485	19.0	1:06	800

首先对样本进行主成分分析，我们选择x_2，x_3，x_4，x_5，x_6，x_7，x_8为主成分分析的对象，分析结果如下：

表 3.3: 特征值和主成分贡献率

	Component	Total	%of Variance	Cumulative%
Raw	1	97.109	81.565	81.565
	2	8.720	7.324	88.890
	3	5.753	4.832	93.722
	4	3.953	3.320	97.042
	5	2.022	1.698	98.740
	6	0.895	0.752	99.492
	7	0.605	0.508	100.000

如表3.3所示，Total列分别为7个主成分的特征值，前4个主成分的累积贡献率为97.042，选前4个主成分就可以。

可以根据表3.4矩阵给出的系数，计算主成分分析后的样本数据，表3.5为主成分分析后的样本数据。

表 3.4: 主成分得分的系数矩阵

	Component			
	1	2	3	4
x_2	-0.834	0.578	1.288	0.537
x_3	0.001	-0.003	0.028	0.125
x_4	0.182	0.546	1.588	0.097
x_5	0.019	-0.224	0.038	0.956
x_6	0.002	0.016	0.030	-0.218
x_7	-0.004	0.019	-0.006	-0.004
x_8	0.020	0.834	-0.487	0.402

表 3.5: 主成分分析后的样本数据

02/16/1998	-2.37626	0.62633	-1.12140	0.15068
04/15/1998	2.17939	-0.25617	-0.87126	-0.21770
04/19/1998	-1.73315	-0.69199	1.88497	-1.31181
04/27/1998	-1.66903	-0.89196	-0.32908	0.05433
04/30/1998	0.68828	-0.96754	-0.72023	-0.49967
05/20/1998	-0.08675	-1.12315	0.04697	-0.38235
11/01/1998	1.69509	0.57068	0.96534	0.10146
02/17/1999	0.20784	1.17913	1.26304	-0.21512
03/15/1999	-1.19584	0.28733	-1.80350	0.45822
04/04/1999	0.68971	0.26048	0.05024	-1.14660
04/10/1999	0.91241	-0.49392	-0.94493	-0.10847
07/10/1999	0.90431	-1.29417	-0.58831	0.28217
08/05/1999	-1.19049	-2.21584	0.61059	-0.75599
11/23/1999	0.81512	0.81882	-0.76309	-0.14837
03/21/2000	-1.49424	0.22356	-1.16212	0.44091
03/26/2000	0.74807	0.60695	0.97798	-0.69947
04/09/2000	0.07273	0.12472	0.93400	-1.06268
04/12/2000	1.11762	0.01002	-0.43886	2.93440
06/05/2000	0.57567	-1.35151	-0.45836	-0.16251
07/06/2000	0.29091	-1.56317	0.34711	-0.76318
08/01/2000	-0.03289	-1.43068	0.83356	1.44239
10/17/2000	-0.18072	0.49849	2.67004	0.37500
11/18/2000	0.15600	1.23214	0.62140	0.45348
12/23/2000	0.19534	1.27047	0.04205	-0.68848
12/31/2000	0.34934	1.34626	-0.38159	-0.25425
01/02/2001	-0.38022	1.34115	-0.28759	-1.39863
01/12/2001	-0.94657	1.32951	-1.91932	-0.56181
03/04/2001	-0.72216	1.08626	1.03308	1.65701
03/19/2001	0.75564	0.62802	-0.30146	-1.49867
04/06/2001	-0.55490	-0.17263	0.20637	1.07461
04/08/2001	0.64095	-0.13206	-1.00093	0.71557
04/09/2001	-0.18139	0.13079	0.19276	-0.71094
04/12/2001	-0.42635	0.52392	0.78649	2.25228
06/11/2001	0.17656	-1.51022	-0.37398	0.19418

图 3.2: LIBSVM 对表3.5数据调试参数C，γ图

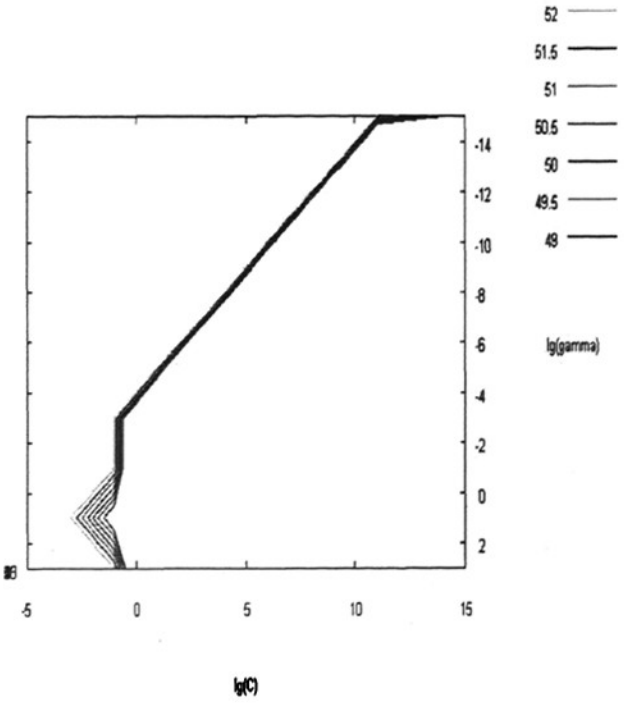
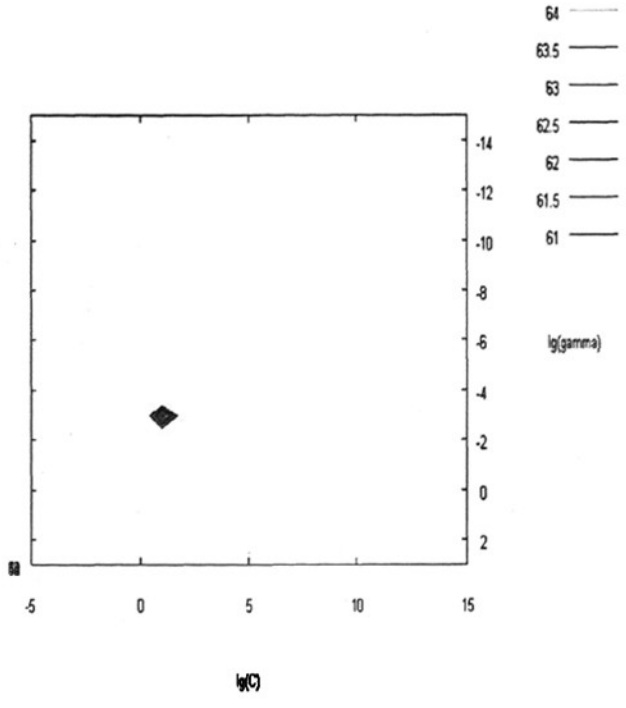


图 3.3: LIBSVM 对表3.2数据调试参数C，γ图



我们根据表3.1把表3.2中按x_{10}能见度划分为三类：黄，橙，红。为应用方便，我们把黄用1表示，橙用2表示，红用3表示。最后我们来用LIBSVM软件包对表3.5做支持向量机分类测试，我们为验证被主成分分析特征提取后的支持向量机判别效果比不进行处理得好，对表3.2中的原始数据直接用LIBSVM软件包直接处理。经过不断的试验我们最终选用$C-SVC$，径向基核函数，我们用LIBSVM软件包对惩罚因子C和参数γ进行调试，如图3.2，3.3所示的交叉验证精度的等高线图。

对表3.5数据调试出来的最佳参数$C=0.03125$ $\gamma=0.0078125$，调用最佳参数$C=0.03125$ $\gamma=0.0078125$ 对表3.5数据训练结果：

$$optimization\ finished, \#iter=11$$

$$obj=-0.687436$$

$$rho=-0.999129$$

$$Total\ nSV=23$$

调用上述训练结果对表3.5数据预测，结果如下：

$$Accuracy=66.6666\%(6/9)(classification)$$

其中，$\#iter$为迭代次数，obj为SVM文件转换为二次规划求解得到的最小值，rho为判决函数的常数项b，$Total\ nSV$为持向量总个数。

对表3.2数据调试出来的最佳参数$C=2.0$ $\gamma=0.125$ 调用最佳参数$C=2.0$ $\gamma=0.125$对表3.2数据训练结果：

$$optimization\ finished, \#iter=28$$

$$obj=-35.204398$$

$$rho=-0.590663$$

$$Total\ nSV=24$$

调用上述训练结果对表3.2数据预测，结果如下：

$$Accuracy=44.4444\%(4/9)(classification)$$

表 3.6: 两种方法对比分析表

分析方法	样本总数	训练样本数	迭代次数	预测样本数	准确率
主成分分析和支持向量机组合判别	34	25	11	9	66.6666% (6/9)
支持向量机方法(不经过主成分分析变换)	34	25	28	9	44.4444% (4/9)

为了更好地说明基于主成分分析和支持向量机组合判别分析方法判别性能好于没有经过主成分分析变换的支持向量机方法，我们给出了表3.6。

由以上实验结果我们可以看出，经由主成分分析方法提取特征后，支持向量机判别方法的识别准确率得到了大幅度的提升，而且迭代的次数也从28减少到11次，这充分说明了主成分分析方法的有效性，以及特征提取在支持向量机判别方法中的必要性。

这一判别分析方法应用到沙尘暴预警上仍然不能达到非常准确的程度，究其原因，其一是不能够保证我们所获得的沙尘暴资料数据是真实的；其二是我们选取的沙尘暴指标是否已经全面、完整地概括了沙尘暴起因仍是一个疑问；其三我们获得到的数据资料样本数太少影响训练的充分性。其四是不同地方沙尘暴类型有着不同的特点，试图找出一个可以判别所有沙尘暴级别的数学模型是非常困难的。但是如果我们能较正确的较多的获得某一地区的历史数据资料，考虑用此种办法去判别该地区的沙尘暴级别，估计能非常准确作该地区的沙尘暴预警，因为虽然不同地方沙尘暴起因不同，但就某个地区来说，我们能很好的找出影响沙尘天气的主要因素，这种状况有利于我们用主成分分析进行特征提取提高主成分的累积贡献率。另外这种方法也可用于其它领域的预警。

第四章　结论与建议

本文在收集了大量有关主成分分析，支持向量机等资料的基础上，结合主成分分析方法对指标繁多的事物进行特征提取，再结合支持向量机分类方法的特点（通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能），提出了基于主成分分析和支持向量机的组合判别分析方法，来对其事物所属种类进行判别分析，并给出了详细的数据分析操作步骤。通过这种方法，既能简化运算，又可以克服线性方法的局限性，从而提高判断的准确率。本文针对支持向量机分类的核函数选择问题，提出了新的方法—基于粒子群优化算法的组合核函数分类方法。最后将组合判别分析方法应用在沙尘暴预警领域，实验证明判断的准确率确实是有显著的提高。主要结论如下：

1 系统讨论了判别分析方法的理论，研究进展及其方法。

2 针对支持向量机分类的核函数选择问题，提出了基于粒子群优化算法的组合核函数分类方法。

3 结合主成分分析和支持向量机分类理论，给出了基于主成分分析和支持向量机的组合判别分析方法及其数据分析步骤。

4 将组合判别分析方法应用于沙尘暴的预警领域。本文是利用气象数值预报给出的气象指标，结合当地沙尘暴历史数据指标进行判别分析是哪种程度的沙尘暴，也即是说判别分析当前天气指标做出红或黄或橙的预警信号。并给出了一个判别分析流程粗框图，从流程图中我们能清晰的看到基于主成分分析和支持向量机的组合判别分析方法应用于沙尘暴预警的全过程和步骤。

5 结合实际的例子—民勤地区发生34次沙尘暴的前期气象指标数据，做了具体的数据分析试验。试验结果表明：基于主成分分析和支持向量机组合判别分析方法的准确率(66.6666%)明显高于不做主成分分析的支持向量机方法(44.4444%)。而且迭代的次数从28减少到11次，这大大的简化了运算。

6 针对主成分分析和支持向量机的组合判别分析方法在实际例子中分析出的结果，我们得出要能较正确的较多的获得某一事物的历史数据资料，充分挖掘数据的有效性信息，来提高判断准确率。

尽管研究工作在理论和实际应用中取得了一些成果，但论文中也存在着一些问

题与不足，还有许多问题有待研究，下面几点也许值得注意：

1 数据收集问题，基于主成分分析和支持向量机的组合判别分析方法在应用沙尘暴预警时，由于没做深入的研究，时间的不足，信息渠道的不畅，数据收集的不充分，因此准确率并不是非常高，建议样本数据最好有百条以上。

2 数据处理问题，缺失值的处理采用的是删除含有缺失值的个案，但是这样很容易产生信息丢失的问题。本文也没进行异常值的处理，影响了结果的准确性。

3 由于没有不发生沙尘暴的天气气象指标，所以忽略了对非沙尘天气的预测判别。

参考文献

[1] 张润楚，多元统计分析，科学出版社，2006.

[2] 袁志发，周静芋，多元统计分析，科学出版社，2006.

[3] 冯杰，黄力伟，王勤，尹成义，数学建模原理与案例，科学出版社,2008.

[4] Berry, Michael W. Martin, Dian I.Principal component analysis for information retrieval, Handbook of parallel computing and statistics, Stat. Textb. Monogr, 184, Chapman Hall/CRC, Boca Raton, FL,9(2006)399-413.

[5] Huckemann, Stephan,Ziezold. Herbert Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces, Adv. in Appl. Probab. 38 (2006), no.2, 299-319.

[6] Caussinus, Henri, Ruiz-Gazen. Anne Classification and generalized principal component analysis. Selected contributions in data analysis and classification,Stud. Classification Data Anal. Knowledge Organ., Springer, Berlin, (2007).539-548.

[7] Zuccolotto. Paola Principal components of sample estimates: an approach through symbolic data analysis. Stat. Methods Appl. 16 (2007), no.2,173-192.

[8] Sato, Manabu, Ito. Masaaki Theoretical justification of decision rules for the number of factors: principal component analysis as a substitute for factor analysis in one-factor cases. J. Japan Statist. Soc. 37 (2007), no. 2, 175-190.

[9] Fung, Wing K, Gu, Hong, Xiang Liming, Yau Kelvin . Assessing local influence in principal component analysis with application to haematology study data. Stat. Med. 26(2007), no. 13, 2730-2744.

[10] Chuvashova. Fan of the principal component of the toric Hilbert scheme. (Russian) Uspekhi Mat. Nauk 62(2007), no.5.

[11] Hoare, Zo.Landscapes of naive Bayes classifiers. PAA Pattern Anal. Appl. 11 (2008), no. 1, 59-72.

[12] Zhang Ming Wei, Wang Bo, Zhang Bin, Zhu Zhi Liang .Weighted naive Bayes classification algorithm based on correlation coefficients. (Chinese) J. Northeast. Univ. Nat. Sci. 29 (2008), no. 7, 952-955. 167–168.

[13] 王熙照，模糊测度和模糊积分在分类技术中的应用，科学出版社，2008.

[14] 陈晶，肖丁，决策树算法在数据挖掘中的应用研究，软件导刊，Software Guide,2008年03期，98-99.

[15] Kirkos, Efstathios,Spathis, Charalambos Manolopoulos. Yannis Support vector machines, decision trees and neural networks for auditor selection. J. Comput. Methods Sci. Eng. 8 (2008), no. 3, 213-224.

[16] Polat, Kemal,Güne.Salih Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Appl. Math. Comput. 187 (2007), no. 2, 1017-1026.

[17] http://www.gotoread.com/vo/990/page96267.html.

[18] 李雄飞，谢忠时，李晓堂，李军，基于粗集理论的约简算法，吉林大学学报(工学版). Vol. 33 No. 1, Jan. (2003)82-87.

[19] Liu Guilong, Zhu William. The algebraic structures of generalized rough set theory. Inform. Sci. 178 (2008), no. 21, 4105-4113.

[20] 耿子林，权光日，叶风，一种基于扩张矩阵理论的规则学习算法，小型微型计算机系统，Vol. 18 No.6, 46-50.

[21] Chen Zhen Zhou,Zou Li Shan.KNN algorithm based on features extracted by KFST. (Chinese) J. South China Normal Univ. Natur. Sci. Ed. (2008), no. 2, 50-55.

[22] A comparative study of three artificial neural networks for the detection and classification of gear faults. Int. J. Gen. Syst. 34 (2005), no. 3, 261-277.

[23] Classification of dive profiles: a comparison of statistical clustering techniques and unsupervised artificial neural networks. J. Agric. Biol. Environ. Stat. 3 (1998), no. 4, 383-404.

[24] Bandyopadhyay Sanghamitra, Pal Sankar K. Classification and learning using genetic algorithms. Applications in bioinformatics and web intelligence. Natural Computing Series. Springer, Berlin, 2007.

[25] Hu Yi-Chung. Fuzzy integral-based perceptron for two-class pattern classification problems. Inform. Sci. 177 (2007), no. 7, 1673-1686.

[26] 刘顺忠，数理统计理论方法应用和软件计算，华中科技大学出版社，2005.

[27] 王雪，测试智能信息处理，清华大学出版社，2008.

[28] 邓乃扬，田英杰，数据挖掘中的新方法支持向量机，科学出版社，2006.

[29] Vapnik V N .统计学习理论本质，张学工译，清华大学出版社，2000.

[30] Tom M. Mitchell.机器学习[M]. 北京: 机械工业出版社,2005.

[31] 蔡自兴，徐光祐，人工智能及其应用[M]，北京:清华大学出版社, 2000.

[32] 徐从富，李石坚，王金龙.机器学习研究与应用新进展，浙江大学人工智能研究所，2006年10月16日第二稿.

[33] Ancona Netal.Ball detection in static images with Support Vector Machines for classification. Image and Vision Computing. 2003(21),675-692

[34] Vapnik V N.Statistical Learning Theory.John Wiley Sons,1998.

[35] Vapnik V N.The Nature of Statistical Learning Theory.New York: Springer-Verlag,1995.

[36] Burges C J C.A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery ,(1998)21-167.

[37] Gun s. Support vector machines for classification and regression. Technical report,ISIS grop,University of SouthAMPTON,1998.

[38] T. B. Trafalis,R. C. Gilbert.Robust support vector machines for classification and computational issues. Optimization Methods and Software, Volume 22, Issue 1 February (2007)187 - 198.

[39] Sami Ekici. Classification of power system disturbances using support vector machines.Expert Systems with Applications, Volume 36, Issue 6, August (2009) 9859-9868.

[40] Hui Li, Jie Sun. Predicting business failure using multiple case-based combined with support vector machine reasoning. Expert Systems with Applications, Volume 36, Issue 6, August (2009)10085-10096.

[41] Li Chang Chao, Lee Ing Tong. Wafer defect pattern recognition by multi-class support vector machines by using a novel defect cluster index.Expert Systems with Applications, Volume 36, Issue 6, August (2009)10158-10167.

[42] Haritha Saranga. Identifying winners of competitive events: A SVM-based classification model for horserace prediction.European Journal of Operational Research, Volume 196, Issue 2, 16 July 2009, Pages 707-718.

[43] Chenn-Jung Huang, Yi-Ju Yang ,Dian-Xiu Yang, You-Jia Chen. Frog classification using machine learning techniques.Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 3737-3743.

[44] 官理，祖峰，唐文胜，快速的支持向量机多类分类研究，计算机工程与应用，2008, 44(5),177-179.

[45] Bates J M,Granger C W J,The Combination of Corecasts[J],Operational Reserch Quarterly,(1969),451-468.

[46] 陈华友，组合预测方法有效性理论及其应用，科学出版社，2007.

[47] 张冰，孔锐，一种支持向量机的组合核函数，计算机应用，Vol.27, No.1,(2007), 44-46

[48] 黄友锐，智能优化算法及其应用，国防工业出版社，2008.

[49] Kennedy J, Eberhart R C. Particle Swarm Optimization[C]. Proceeding of IEEE Intrmational Conference on Neural Networks Perth, (1995) 1942-1948.

[50] Parsopoulos K E， Vrahatis M N. Recent approach to global optimization problem through particle swarm optimization[J]. Natural Computing,(2002), 235-256.

[51] http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm/

[52] 刘伟东，程从兰，张明英，张小玲，王迎春，北京地区沙尘天气监测预报预警业务系统，气象科技，Vol. 32 ,Suppl,Dec,(2004),50-53.

[53] 常兆丰，梁从虎，韩福贵，马中华，民勤沙区沙尘暴的分布特征及前期特征研究，干旱区资源与环境，Vol.16,No.2 ,2002 年6 月,107-111.

作者读硕士期间的工作

1.An Effective Method of Preventing Power Grid Collapse: Peak Load Analysis and Forecasting, 2nd China Workshop on Information System for Crisis Response and Management & Post-Conference Meeting to the International Disaster Reduction Conference, pp: 551-558 (ISTP收录)

2.Ontology-Based Assembly Design and Information Sharing for Supply Chain Information, The 38th International Conference on Computers and Industrial Engineering, pp: 1220-1226 (ISTP收录)

致 谢

在此论文完成之际，首先衷心感谢我的导师王建州副教授三年来对我的学习和生活给予的悉心指导和无微不至的关怀。导师不仅传授给我知识，他那敏锐的学术洞察、严谨的治学态度、积极进取和刻苦钻研的精神、以及做人的准则，将会激励我在今后的学习和工作中，使我终身受益。

其次，我要感谢数学与统计学院的各位领导、老师和同学们，感谢你们对我的教育、关怀和帮助，母校严谨求实的学术氛围，良好的育人环境时刻激励着我，诸多学识渊博的名师熏陶了我，同学们踏实勤奋、勇于拼搏的精神感染了我，使我有了前进的方向和动力，谢谢你们！此外，要感谢我的同门师弟、师妹们，感谢你们在学习上、生活上给予我的关心与帮助。特别感谢同门朱素玲、梁洪锁对我论文修改上的帮助。有缘与你们成为同门，彼此关怀照顾，谢谢你们，祝愿你们能在以后的学习中更加顺利，更加成熟！最后，我要感谢我的爸爸、妈妈对我的关心和支持，没有他们，就没有今天的我。你们的理解和支持是我坚实的后盾，是你们给了我巨大的物质支持和精神支持，才使我顺利完成学业，谢谢你们！

在我的学业及毕业论文的完成过程中，浸透着很多人的心血和汗水，我在这里再一次向他们表示诚挚的谢意，并祝他们身体健康，一生平安！

Thank You!

王 惠 婷

2009年 5月于兰州大学