

一种有效的大规模数据的分类方法

张艳宁¹, 赵荣椿¹, 梁 怡²

(1. 西北工业大学 756 信箱, 陕西西安 710072; 2. 香港沙田香港中文大学地理系, 香港)

摘 要: 本文提出了一种基于自组织特征映射神经网络(SOM)和支撑矢量机(SVM)相结合的复杂模式的大规模数据的分类方法. 该方法首先利用自组织特征映射神经网络对待识目标进行聚类, 然后应用支撑矢量机方法对其进行分类识别. 通过对复杂异或(XOR)分类问题, 以及实际的 Iris 和 Appendicitis 数据分类问题等的分类实验, 且与仅用支撑矢量机的分类方法比较, 结果表明, 本文提出的方法对复杂模式的大规模数据的分类识别问题具有较好的效果, 且训练时间大幅度减小.

关键词: 自组织特征映射神经网络; 支撑矢量机; 大规模数据; 模式分类

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2002) 10-1533-03

An Efficient Target Recognition Method for Large Scale Data

ZHANG Yan-ning¹, ZHAO Rong-chun¹, LEUNG Yee²

(1. P. O. Box 756, Northwestern Polytechnical University, Xi'an, Shanxi 710072, China;

2. Department of Geography, The Chinese University of Hong Kong, Shatin, Hong Kong, China)

Abstract: An efficient target recognition method for large scale data is proposed in this paper, which is based on self-organizing map (SOM) neural network and support vector machines (SVMs). The target data set is divided into clusters by SOM first. Then, the support vector machines are applied to classify targets. The new method is used to classify the complex XOR problem, Iris and Appendicitis data, and the experimental results show that the new method can obtain better recognition results for the complex pattern classification of large scale data, and the training time is shorter than that by using the support vector machine method only.

Key words: self-organizing map neural network; support vector machines; large scale data; pattern classification

1 引言

目前, 随着计算机、网络以及通讯技术的发展, 针对大规模数据的模式分类识别问题日益受到了人们的重视. 模式分类的方法有很多, 例如: 统计模式分类法、结构法、分类树以及神经网络等, 但对于高维和海量分类识别问题, 以及没有先验知识的情况下, 这些方法都难以得到满意的结果^[1]. 传统的神经网络方法由于对数据的过拟合, 而导致其推广能力的下降. 如果克服了过拟合的缺点, 神经网络的推广能力将会得到大大提高. 最近, 一些应用表明支撑矢量机方法显示出较传统方法包括神经网络方法更好的推广能力^[2,3]. 支撑矢量机是由 Vapnik 和他的研究组提出的一种新的模式识别技术^[2,4,5]. 它来源于在两类中具有最大区间的最优越平面具有最好推广能力的思想. 在支撑矢量机中输入矢量被非线性地映射到高维特征空间中, 在新的空间中, 一个线性决策面的建立将导致原始输入空间中一个非线性决策面的产生. 在研究中我们发现, 对于复杂高维大规模模式识别问题, 支撑矢量机方法的训练时间较长. 针对这一问题, 基于降低问题复杂度的考虑, 我们提出了基于自组织特征映射神经网络(SOM)和支撑矢量机

(SVM)方法相结合的模式分类识别方法.

2 自组织特征映射神经网络和支撑矢量机

2.1 自组织特征映射神经网络(SOM)

SOM 是一种基于非监督和竞争学习的人工神经网络模型^[3], 它可实现高维输入空间到二维格状神经元空间的映射, 其中每一个神经元可代表一种模式分布. 因此, SOM 可用于复杂的高维数据聚类分析.

SOM 网络结构为一单层的前馈网络, 包含一输入和激活层. 激活层由一些神经元在一二维平面上排列组成. SOM 具有训练和分类两种状态. 当输入矢量与某个连接输出神经元和输入神经元的权值的欧氏距离最小时, 该输出神经元被激活, 并作为网络的输出, 此时该连接权被修正为和输入矢量更接近, 而该输出神经元也被称为竞争获胜神经元. 相应邻域连接权值也得到修改, 直至网络达到终止条件. 邻域函数、学习速率和终止条件都随具体问题而定. SOM 的训练步骤参见参考文献^[3].

收稿日期: 2001-08-17; 修回日期: 2002-02-17

基金项目: 航空基础研究基金(No. 00F53050); 雷达信号处理重点实验室基金(No. 2000JS01. 4. 1. HK0311); 陕西省自然科学基金(No. 2001cs1102)

2.2 支撑矢量机(SVM)

由于大多实际模式识别问题是非线性可分的,这里只介绍解决非线性可分问题的支撑矢量机方法.考虑一个两类模式的识别问题,假设训练集为: $\{(X_i, d_i)\}_{i=1}^N$, $X_i \in \mathbf{R}^m$ 是对应与第 i 个样本的输入模式, $d_i \in \{-1, +1\}$ 是对应与第 i 个样本的期望输出, N 为训练样本数.非线性支撑矢量机首先进行非线性映射 $\phi: \mathbf{R}^m \rightarrow \mathbf{R}^{m_1}$, 设 $\{\phi_j(X)\}_{j=1}^{m_1}$ 为输入空间 $\mathbf{R}^m$ 到输出空间 $\mathbf{R}^{m_1}$ 的一组非线性变换,则该映射可由下式表示:

$$\Phi(X) = [\phi_1(X), \phi_2(X), \dots, \phi_{m_1}(X)]^T \quad (1)$$

一个支撑矢量机可以训练来构成一个超平面 $w^T \Phi(X) + b = 0$ 使得其分离区间最大^[4]. 式中 w, b 为可通过训练求得. 应用 Lagrange 乘子法, 该超平面可被表示成:

$$\sum_{i=1}^N \alpha_i d_i \Phi^T(X_i) \Phi(X) = 0 \quad (2)$$

这里变量 α_i 是 Lagrange 乘子. 这些 α_i 可从求解下列问题得到^[4]:

$$\text{maximize: } Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j d_i d_j K(X_i, X_j) \quad (3)$$

满足约束: $\sum_{i=1}^N \alpha_i d_i = 0, 0 \leq \alpha_i \leq C$ 且 $i = 1, 2, \dots, N$. 这里 C 是一个常数. 该参数值可经验地取 10. $K(X_i, X_j)$ 是核内积, 定义为:

$$K(X_i, X_j) = \Phi^T(X_i) \Phi(X_j) \quad (4)$$

这里我们应用一个高斯核函数定义为:

$$K(x, y) = \exp\left[-\frac{(x-y)^2}{2\sigma^2}\right] \quad (5)$$

从 Kuhn-Tucker 条件^[4]可以看出, 仅仅位于决策边界的训练样本, 相应的 Lagrange 乘子才为非零. 可以看出, 只有这些被称作支撑矢量的样本, 才影响着超平面的建立.

因此, 决策函数可由下式获得^[4]:

$$f(x) = \text{sgn}\left[\sum_{i=1}^N \alpha_i d_i K(X_i, x) + b\right] \quad (6)$$

3 基于 SOM 和 SVM 的模式分类识别方法

在利用 SVMs 对实际复杂大规模数据进行分类识别研究时, 其训练时间较长, 当复杂度达到较高程度时, 其所需训练时间将给实际应用带来很大困难. 如果能适当降低问题复杂度, 则这一问题可得到一定程度的解决.

为了分析简单起见, 这里以二类识别问题为例, 多类识别问题可以此类推. 一、二类识别问题: 特征矢量维数为 m , 类别数为 2, 则可描述为寻找一映射: $f: \mathbf{R}^m \rightarrow \mathbf{R}^2, Y = f(X)$, 其中 $X \in \mathbf{R}^m, Y \in \mathbf{R}^2$. X 可被划分为空间子集: $X = X_1 \cup X_2 \cup \dots \cup X_N$, 其中 $X_i \cap X_j = \emptyset, j = 1, 2, \dots, N$; 因此, Y 可被描述为子集的集合: $Y = Y_1 \cup Y_2 \cup \dots \cup Y_N$, 则映射 f 可分解为一个映射组:

$$\begin{cases} f_1: \mathbf{R}^m \rightarrow \mathbf{R}^2 & Y_1 = f_1(X_1) \\ f_2: \mathbf{R}^m \rightarrow \mathbf{R}^2 & Y_2 = f_2(X_2) \\ \dots\dots\dots \\ f_N: \mathbf{R}^m \rightarrow \mathbf{R}^2 & Y_N = f_N(X_N) \end{cases} \quad (7)$$

这里 $f_i, i = 1, 2, \dots, n$ 表示一个二类识别映射. 分解后映射组中每一 f_i 的复杂度远小于分解前映射 f . 显然, 寻找映射 f 就等于寻找映射 $f_i, i = 1, 2, \dots, N$. 该方法包含两个步骤: 将数据集划分为空间子集, 以及实现子映射 f_i .

该方法的思想是, 先用某种方法对数据集进行初分类, 将其划分为空间子集, 然后再利用某种分类识别方法对各子集进行分类识别.

通过第二节和第三节分析, 我们知道: SOM 可对数据集进行聚类分析, 其结果可获得数据集的各种模式的子集; SVMs 是较传统模式识别方法具有更好的适应和推广能力. 基于以上思想和分析, 本文提出了基于 SOM 和 SVMs 相结合的大规模数据目标识别方法. 该方法的思想是: 在对目标进行特征提取, 形成特征矢量数据集后, 利用 SOM 方法对数据集进行划分, 即对目标特征矢量数据集进行聚类, 从而将目标特征空间划分为有一定聚类中心的子集, 子集数可根据具体问题初步估计. 也就是说, 首先将目标特征空间的复杂度适当降低, 然后应用 SVMs 实现子映射, 即对各子集再作进一步的分类识别, 这样既可以减少 SVM 的训练时间, 又可以保持其识别效果好的优点.

4 仿真实验和实际数据分类识别

本节中我们将利用本文提出的方法和支撑矢量机方法, 针对大规模数据复杂异或分类识别计算机仿真数据, 以及实际的分类识别数据 (Iris 和 Appenicitis 数据) 进行分类识别实验. 在实际分类识别实验中, 在训练时, 我们在数据集中均匀抽取训练数据, 即可保证训练数据中的各子集和检验数据中的各子集分布相同.

4.1 复杂异或分类问题仿真实验

实际复杂的模式分类问题往往含有如下特点: 非线性可分和不同类别数据混叠在一起. 考虑到这一特点, 我们用计算机随机产生一个两类复杂异或数据集, 数据分布如图 1 所示. 该数据集由 25 个小区域组成, 每个小区域共含数据 100 个, 单纯的第一类和第二类数据分别共有 5 和 4 个小区域, 两类数据混叠区域共有 16 个. 该数据集包含两类, 2 个特征, 每类数据分别包括 1300 和 1200 个数据. 我们将数据集分为两组, 一组用于训练, 一组用于检验. 训练集两类数据个数分别为 650 和 600, 检验集中两类数据个数分别为 650 和 600.



图1 复杂异或两类分类问题分布图示

我们利用本文提出的方法和支撑矢量机方法分别对两组数据进行训练和检验, 检验结果见表 1. 其中自

组织特征映射神经网络结构为: 输入层 2 个神经元, 输出层 4 × 5 个神经元.

表1 复杂 XOR 检验集识别结果

	SVM 方法	本文方法
正确识别率 (%)	64.96	66.96
训练时间 (秒)	116314.8	82

从表 1 可以看出: 基于自组织特征映射神经网络和支撑

矢量机相结合的方法是解决复杂大规模数据目标识别问题的一个有效方法,该方法较单独使用支撑矢量机方法的训练时间大大减小。

4.2 实际数据分类识别实验

这里,我们将本文提出的方法应用于著名的 Iris 和 Appendicitis 数据分类识别问题。

数据特征性如下:

Iris 数据:3 类,4 个特征,每类 50 个数据;

Appendicitis 数据:2 类,7 个特征,106 个数据(其中第一类 85 个数据,第一、二类 21 个数据)。

我们利用本文提出的方法和支撑矢量机方法分别对两种数据进行训练和检验,检验结果见表 2。其中自组织特征映射神经网络结构为:输入层神经元数与其特征矢量维数相同,输出层 3×4 个神经元。

从表 1 和表 2 可以看出:基于自组织特征映射神经网络和支撑矢量机相结合的方法是解决复杂大规模数据目标识别问题的一个有效方法,该方法较单独使用支撑矢量机方法的训练时间大大减小,且识别结果较好。

表 2 Iris 和 Appendicitis 数据实验结果

	SVM 方法		本文方法	
	Iris	Appendicitis	Iris	Appendicitis
错误识别率 (%)	4.0	25.18	2.0	11.18
训练时间(秒)	56.2	68.3	8.3	12.1

5 结论

本文提出了一种基于自组织特征映射神经网络和支撑矢量机相结合的用于大规模数据的复杂模式的分类方法。该方法首先利用自组织特征映射神经网络对待识目标进行聚类,然后应用支撑矢量机方法对其进行分类识别。通过对复杂异或分类问题,以及实际的 Iris 和 Appendicitis 数据分类问题的分类识别实验,与仅用支撑矢量机的方法进行比较,结果表明,这一方法对大规模数据的复杂模式分类问题具有较好的分类识别效果,且训练时间较单独使用支撑矢量机方法大大减小。

参考文献:

- [1] Gabisch M, Nicolas J M. Classification by fuzzy integral: performance and tests [J]. Fuzzy Sets and Systems, 1994, 65: 255 - 271.
- [2] Vapnik V N. The Nature of Statistical Learning Theory [M]. NY: Springer-Verlag, 1995.
- [3] Kohonen T. Self-organizing Maps [M]. NY: Springer-Verlag, 1995.
- [4] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20: 273 - 297.
- [5] Scholkopf B, Sung K, Burges C, Girosi F, Niyogi P, Poggio T, Vapnik V. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers [R]. A. I. Memo 1559, MIT, 1996.

作者简介:



张艳宁 女, 1967 年出生, 博士, 教授, 西北工业大学计算机系语音与图像教研室主任, 兼任中国电子学会信号处理分会委员, 中国体视学会图像分析分会秘书长, 陕西省信号处理学会秘书长, 中国电子学会高级会员, IEEE 会员, 主要从事信号与图像处理、模式识别、计算机视觉、数据挖掘和智能信息处理等研究, 获省部级科技进步奖 4 项, 陕西省青年科技奖, 在国内外发表学术论文三十余篇。



赵荣椿 男, 1937 年 11 月出生, 教授, 1960 年毕业于中国人民解放军军事工程学院, 现任西北工业大学信号处理研究所所长, 博士生导师, 科研成果多次获得国家及部级奖励, 已出版专著 5 部, 发表科技论文 120 余篇。兼任中国体视学会副理事长、中国图象图形学会副理事长、中国电子学会信号处理分会副主任委员等, 其研究领域包括信号处理、计算机视觉、模式识别等。

梁 怡 男, 教授, 香港中文大学地理系讲座教授及系主任, 获中文大学学士及美国科罗拉多大学文学硕士, 理学硕士及博士学位。担任多种国际杂志编委和学会成员, 研究方向为空间分析、环境工程与专家系统、地理信息系统及决策支持系统、模糊逻辑与神经网络、遗传算法等。出版专著 3 部, 在国际学术刊物和国际会议上发表论文 100 多篇。