

基于对应分析的支持向量机分类研究

王娟, 贺兴时, 赵飞军

(西安工程大学理学院, 西安 710048)

摘要: 提出了基于对应分析的支持向量机分类模型。该模型通过对应分析可以同时变量及样本进行降维和消除相关性, 从而在降低 SVM 训练时间的基础上有效地提高了 SVM 的分类精度。实验结果表明该方法是可行的。

关键词: 支持向量机; 对应分析; 分类模型; 因子分析

中图分类号: TP183

文献标识码: A

引言

对应分析^[1]是 R 型因子分析和 Q 因子分析的结合, 它也是利用降维的思想达到简化数据结构的目的, 对应分析与因子分析不同, 它不但可以对数据表中的行与列进行处理, 而且能以低维图形表示数据表中行与列之间的关系。

对应分析法是在 R 型和 Q 型因子分析的基础上发展起来的一种多元统计分析方法, 因此对应分析又称为 R-Q 型因子分析, 这是由法国 J. P. Benzeci 于 1970 年首次提出, 属于广义因子分析中的一个分支。对应分析克服了 R 型和 Q 型分析没有考虑到 (特征) 变量之间的相关关系对训练样本选择的影响。它综合了 R 型和 Q 型因子分析的优点, 并将它们统一起来使得由 R 型的分析结果很容易得到 Q 型的分析结果, 这就克服了 Q 型分析计算量大的困难。它的另外一个特点是能把众多的样本和众多的变量同时反映到同一个二维平面图上求解, 将样本的属性及其 (特征) 变量在图上直观地表示出来。目前, 该方法在医药卫生领域^[2]和生态环境领域^[3]及大气环境质量评价领域^[4], 取得了较好的应用效果。

本文提出基于对应分析的支持向量机分类算法, 主要是利用对应分析能够克服 SVM 算法在训练过程中没有考虑数据变量及样本之间可能存在的相关性, 且统计学习理论是针对小样本情况进行的, 因而 SVM 有时对

数据的分类并不是十分准确, 在大样本情形下还会出现训练速度过慢的缺点^[5], 本文的主要研究在于克服 SVM 算法的不足。实验与分析表明本文提出的方法是有效、可行的。

1 对应分析的步骤

设数据矩阵为 $X = (x_{ij})_{n \times p}$ 列向量表示指标向量, 行向量表示样本向量^[6,7]。

(1) 原始数据规格化。将 X 的每一个元素转化为正数, 如果某个元素是负数, 则对所有元素同时加上一个数, 使每一个元素都为正数。然后, 将 X 的每一个元素都除以全部元素之和, 得矩阵 P

$$P = (P_{ij}) = \frac{1}{x \cdot} (x_{ij}) \quad (1)$$

其中, $x \cdot = \sum_i \sum_j x_{ij}$

(2) 计算矩阵 Z 。矩阵 $Z_{n \times p}$ 的元素为:

$$z_{ij} = \frac{P_{ij} - P_i \cdot P_j}{\sqrt{P_i \cdot P_j}} \quad (2)$$

其中, $P_i \cdot = \sum_j P_{ij}$, $P_j = \sum_i P_{ij}$

(3) 计算 R 型因子分析的因子载荷矩阵。计算 $A_{pp} = Z'Z$ 的特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$, $0 < r \leq \min\{p, n\}$ 相应的特征向量为 e_1, e_2, \dots, e_r , 并使之单位化。取两个最大特征值 (根据实际需要取特征值的个数) 及其特征向

量, 得因子载荷矩阵为:

$$F = (e_1 \sqrt{\lambda_1} \quad e_2 \sqrt{\lambda_2}) \quad (3)$$

(4) 计算 Q 型因子分析的因子载荷矩阵。计算 $B = ZZ'$ 的特征向量为 $v_i = Ze_i$, 并使之单位化。取两个最大特征值及其特征向量, 得因子载荷矩阵为:

$$G = (v_1 \sqrt{\lambda_1} \quad v_2 \sqrt{\lambda_2}) \quad (4)$$

(5) 在因子轴平面上作指标散点图和样本散点图。指标散点图是以矩阵 F 的两列元素为坐标的散点图, 样本散点图是以矩阵 G 的两列元素为坐标的散点图。

2 集成对应分析的 SVM 训练

支持向量机 (support vector machines, 简称 SVM) 是 V. Vapnik 等在 1992-1995 年间提出的基于统计学习理论 (statistical learning theory, 简称 SLT) 的一种新型机器学习方法。SLT 专门研究实际应用中有有限样本情况的机器学习规律理论。机器学习研究从观测数据出发寻找规律, 利用这些规律对未来数据或无法观测的数据进行预测^[8-9]。SVM 这一新的通用学习方法, 由于它基于结构风险最小化 (SRM) 原理, 而不是传统统计学的经验风险最小化 (ERM), 表现出很多优于已有方法的性能, 迅速引起各领域的注意和研究兴趣, 取得了大量的应用研究成果, 推动了各领域的发展。训练 SVM 的本质就是解决一个二次规划问题^[10-11]。

SVM 的基本原理^[12]是在一定条件下, 建立一个决策曲面即最优超平面, 用以划分一组 n 维向量的训练样本 $(x_1, y_1), \dots, (x_b, y_b)$, 其中 $x_i \in R^n$, 对应的期望输出 $y_i \in \{-1, +1\}$ 。距离这个最优超平面最近的异类向量就是所谓的支持向量 (support vector), 支持向量之间的距离最大 (即边缘最大化), 一组支持向量可惟一确定一个超平面, 如图 1 所示。

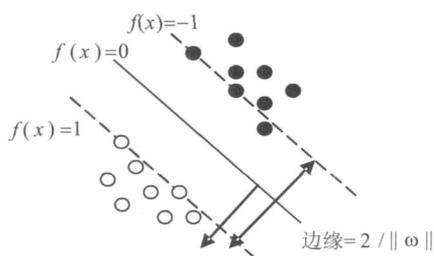


图 1 最优超平面

2.1 训练样本的选择^[7]

- (1) 建立原始的数据矩阵 X ;
- (2) 根据式 (1) 对原始数据矩阵 X 进行规格化;
- (3) 根据式 (2) 计算得到过渡矩阵 Z ;
- (4) 求解 $A_{pp} = Z'Z$ 的特征根 λ_k 和其相应的特征向

量 e_k ;

(5) 由对偶定理得到样本间的协方差矩阵 $B = ZZ'$ 的特征根 λ_k 其相应的特征向量 $v_i = Ze_i$;

(6) 根据 (3) 式和 (4) 式可以得到变量和样本的载荷矩阵 F 和 G ;

(7) 根据样本载荷矩阵 G , 从原始的观测样本中选择出一类的代表性样本集 (序号) $Samples = \{i | \frac{m_i}{n} \geq \alpha | |g_{ij}| \}; j = 1, \dots, m\}$, i 代表原始的观测样本的序号, 而 $|g_{ij}|$ 表示对元素取 $|g_{ij}|$ 绝对值;

(8) 重复上述的 (1) - (7) 选出各类的代表性样本, 作为训练样本进行训练或学习;

(9) 将选出的训练样本作为支持向量机的前置输入系统进行训练。

2.2 实验与分析

实验中所用到的数据是利用 MATLAB 软件随机生成的 2000 个 6 维数据的数据集, 利用聚类分析将其分成三类: 第一类有 605 个样本; 第二类有 642 个样本; 第三类有 753 个样本。

通过对应分析的结果可以得到, 前三个因子的 Cumulative Proportion 是 74.9%, 前四个因子的 Cumulative Proportion 是 92.1%, 即对原始数据的解释程度达到 92.1%。为了直观地显示基于对应分析的样本选择结果在样本的特征空间 (图 2) 中横坐标用特征的第一主分量 (经过主分量变换得到, 便于在平面上显示) 表示, 纵坐标用特征的第二主分量表示。+ 表示原始样本, o 表示经过对应分析选中的典型样本。

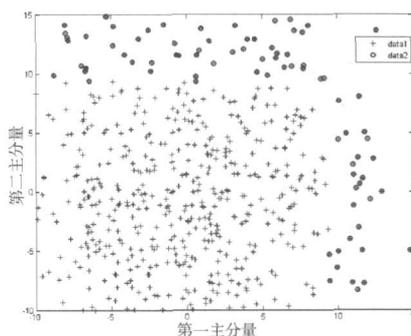


图 2 基于对应分析典型样本的选择结果

从 2000 个数据样本中随机抽取 500 个样本作为测试集, 分别用基于对应分析的 SVM 和 SVM 进行训练与分类测试, 实验结果见表 1。

从表 1 可以得出: 本文提出的方法可以进一步提高训练速度, 因为支持向量机分类的时间主要是由被称为支持向量的那部分样本数决定的, 支持向量越少所用的训练时间就越短。同时基于对应分析的 SVM 方法对测

表 1 集成对应分析的 SVM 与 SVM 算法比较

算法	参与训练 样本数	支持向 量数	训练总 时间 /s	分类正确 率 /%
基于对应分 析的 SVM	900	705	0.335	94.8
SVM	1500	1162	0.562	90.6

试数据集的分类正确率要略高于 SVM 方法,说明基于对应分析的 SVM 方法在分类正确率和训练速度方面都比标准的 SVM 方法有所改善。

参考文献:

- [1] 何晓群. 现代统计分析方法与应用 [M]. 北京: 中国人民大学出版社, 2007.
- [2] 高晓凤. 两种抗结核方案的卫生技术评估及住院结核病人属性特征与治疗转归的对应分析 [D]. 成都: 四川大学, 2007.
- [3] 苏艺, 许兆义, 鄢贵权. 对应分析方法在地下水环境系统分析中的应用 [J]. 北方交通大学学报, 2004, 28(4): 48-53.
- [4] 朱秀华, 李晓旭. 对应分析法在城市大气环境质量评价中的应用 [J]. 大连交通大学学报, 2004, 31(1): 89-94.
- [5] 彭红毅, 蒋春, 杜明. 基于 ICA 与聚类分析的支

- 持向量机分类研究 [J]. 计算机工程与应用, 2008, 44(8): 169-171.
- [6] 虞欣, 郑肇葆. 基于 Q 型因子分析的训练样本的选择 [J]. 测绘学报, 2007, 36(1): 67-71.
- [7] 虞欣. 基于对应分析的训练样本的选择 [J]. 测绘科技情报, 2007, 36(2): 1-7.
- [8] 边肇祺, 张学工. 模式识别 [M]. 北京: 清华大学出版社, 2000.
- [9] Vapnik V, Levin E, Le C Y. Measuring the VC Dimension of a Learning Machine [J]. Neural Computation, 1994(6): 851-876.
- [10] Buşgeç C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.
- [11] 唐小彪. 基于对应分析的支持向量机回归在地震储层厚度预测中的应用 [J]. 物探与化探, 2009, 33(4): 468-471.
- [12] Chapelle O, Haffner P, Vapnik V N. Support vector machines for histogram-based image classification [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1055-1064.

Research of SVM's Classification Based on Correspondence Analysis

WANG Juan, HE Xing-shi, ZHAO Fei-jun

(School of Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract A model of SVM's classification based on correspondence analysis is proposed. This model not only can reduce the dimension of variables and samples but eliminate the correlation between them. Thereby, the classification accuracy of SVM can be effectively improved by reducing the training time of SVM. Experimental results show that the method is feasible.

Key words Support Vector Machine (SVM); correspondence analysis; classification model; factor analysis