多分类支持向量方法的本科招生生源质量模型

邢 涛[1], 廖 冉[2]

(1. 北京航空航天大学 经济管理学院, 北京 100191, xt@buaa.edu.cn;
2. 北京航空航天大学 数学与系统科学学院, 北京 100191)

摘 要: 为了建立一套科学、有效、完善的招生管理流程和评价体系,对多分类支持向量机算法进行改进,利用各地区学生在大学期间的学习成绩与学生入学前的信息特点,探讨本科生源质量预测体系,建立了本科生源质量分析模型.结果表明,改进的多分类支持向量机能够较好地应用于高校本科招生生源质量分析研究,建立的招生生源质量模型可以对本科生源质量进行良好的预测,为高校制定招生计划、提高生源质量提供有效的参考.通过对惩罚矩阵进行合理的调整,模型学习的结果和可靠性都有显著提高.

关键词: 数据挖掘;多分类支持向量机;统计学习理论;招生;质量分析
中图分类号: TP183 **文献标志码:** A **文章编号:** 0367-6234(2010)11-1828-05

Multi-category support vector machine algorithm for quality analysis of student enrollment of universities

XING Tao[1], LIAO Ran[2]

(1. School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China, xt@buaa.edu.cn;
2. School of Mathematics and Systems Science, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: In order to explore and establish a scientific effective and comprehensive recruitment management process and valuation system, and based on several information before students entering college, such as regions and location, the score in college entrance examination, the rank in certain regions, a new model to deal with enrollment strategy and related matters has been set up in this paper. By a real case, it shows that the model is suitable to analyze the quality of fresh students, and can provide proper guidance for planning college enrollment stralegy. The model reliability and result accuracy can be significantly inproved by reasonable adjustment of penalty matrix.

Key words: data mining; multi-category support vector machine; statistical learning theory; students´sources enrollment; analysis of quality

近年来,高等教育已从“精英教育”快速向“大众化教育”转变,高校招生规模逐年扩大,使得高校在生源上竞争日趋激烈.怎样从全国的高考生中录取更多的优质生源,成为各高校招生工作者的一大难题. 各高校招生工作部门一直在积极探索一套科学、有效、完善的招生管理和评价体系[1].在现行招生制度下,高考成绩仍然是绝对性指标.随着高校招生改革的推进,高考自行命题省份不断增加,高考也从最初的全国统一标准变得更具区域性特色.因此,如何在兼顾教育公平的原则下,合理安排招生计划,从而在全国众多的报考学生中科学、公正、高效地挑选出有发展潜力的优秀学生是提高生源质量的关键问题. 以往的高校本科招生生源分析,并未充分挖掘高校历年已录取的学生信息.教育的发展具有持续性,一个地区以往的生源质量往往能够反映出该地区的教育水平,从而指导今后的招生工作,而高校本身拥有的庞大的学生信息资源也恰好利于这一方法的实施.

统计学习理论是专门研究有限样本情形下的

收稿日期: 2009-02-21.
作者简介: 邢 涛(1974—),男,博士研究生.

机器学习规律的理论，在这一理论基础上发展出了一种通用学习算法，相比起传统统计方法所采用的经验风险最小化原理，支持向量机算法采用了结构风险最小化理论并体现出了优于传统统计方法的一些性能[2]. 多分类支持向量机算法是传统支持向量机算法的推广，该算法不仅继承了支持向量机算法稳健且计算精确等特点，而且克服了传统支持向量机算法只能解决二分类问题的局限.

本文利用多分类支持向量机算法，根据各地区学生在大学期间的学习成绩，与学生入学前的信息特点，探讨本科生源质量预测体系，建立本科生源质量分析模型，为高校制定招生计划、提高生源质量提供有效的参考信息. 该模型可以对高校的招生工作给出初步的参考.

1　支持向量机

支持向量机算法是建立在统计学习理论的结构风险最小化原理基础上的，能较好地解决小样本问题，同时具有很好的泛化能力. 算法由Vapnik等[3]在 1995 年左右提出. 近些年来，该算法在实际应用及理论推导中都有着长足的进步. 韩媚和郭丹凤[4]将支持向量机方法应用到了高校就业情况的分析中. 向小东和宋芳[5]利用基于核主成分分析的支持向量机算法对福建省的失业率进行了有效的预测和分析. 相比其他传统的统计学分类方法或是带有学习过程的人工智能算法，如神经网络方法，支持向量机算法能够克服过学习问题和维数灾难问题，具有全局最优的特点和很好的泛化能力[3]. 该算法主要有以下两个特点[6]：

(1)采用结构风险最小化原则，所得到的学习机器有着很好的泛化能力，即由有限的训练样本得到的小误差能够保证对独立的测试集仍保持小的误差.

(2)支持向量机算法最终将转化为一个凸优化问题，局部最优解一定是全局最优解.

支持向量机算法的基本思想是在样本空间或者特征空间，构造出最优超平面，尽可能将不同类的数据点分开，并使得超平面与不同类数据点之间的距离最大，从而达到最大的泛化能力. 在线性可分情况下，如图 1 所示，圆圈和方块分别代表两类样本点，每类中处在边界的点为支持向量，这些支持向量所形成的超平面 H_1 和 H_2 为能撑起不同类别样本点的分类超平面，其之间的间隔为类与类之间的最大间隔，因此最佳的分类面 H_0 介于这两个分类超平面之间.

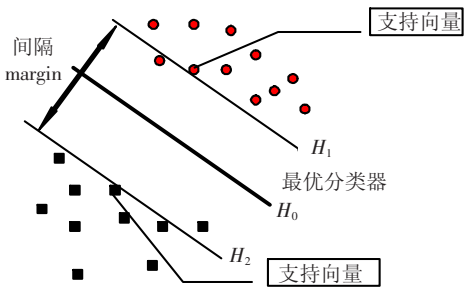

图 1　线性可分下的支持向量机模型

假设样本集合为 $S = \{(\boldsymbol{x}_i, y_i) \in \mathbf{R}^n \times \{\pm 1\}, i = 1, \cdots, m\}$，其中 $\boldsymbol{x}$ 为输入样本，y 为输出样本. 用 $\boldsymbol{x} \in A, y = 1$ 表示一类点，$\boldsymbol{x} \in B, y = -1$ 表示另外一类点. 算法的目标是利用训练样本提供的信息来估计一个分类器，即函数 $f: \mathbf{R}^n \to \{\pm 1\}$. 如果训练样本是线性可分的，那么必然存在 $(w, b) \in \mathbf{R}^n \times \mathbf{R}$ 使得 $w^{\mathrm{T}}x + b = 0$ 为其线性边界，且满足 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \geqslant 1 (\boldsymbol{x} \in A)$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \leqslant -1 (\boldsymbol{x} \in B)$. 该限制条件可被表示为：$y_i[(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}_i) + b - 1] \geqslant 0$，得到决策函数：$f_{w,b}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$，其中 $\boldsymbol{w}$ 为权重向量，b 为偏离值. 最终该问题可以转化成为求解如下二次优化问题：

$$\begin{cases} \min \dfrac{\| \boldsymbol{w} \|^2}{2}, \\ \text{s. t. } y_i[(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}_i) + b - 1] \geqslant 0, \ i = 1, \cdots, m. \end{cases}$$

然而，在实际问题中常常遇到线性不可分的情形，支持向量机算法的优势也体现在其能处理非线性复杂系统中的问题，它能通过非线性变换将该非线性问题转化成为某个高维空间中的线性问题，在高维空间中求取最优分类面实现线性分类. 这种非线性变换是通过定义适当的核函数实现的，在本文中特别应用到了希尔伯特再生核空间 $\boldsymbol{H}_K$[7] 中的两种常用的核函数，其定义如下：

高斯核函数：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{\sigma^2}\right\}, \ \sigma > 0,$$

多项式核函数：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = [(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + 1]^q, \ q \in \mathbf{N}^*.$$

2　多分类支持向量机

支持向量机的产生最初是为了解决二分类问题，但在实际应用中，往往要求算法可以对多于两类的分类问题给出准确的判断. 因此，随着实际应用的广泛要求和理论研究的深入，多分类支持向量机算法就此产生.

多分类支持向量机算法在很多领域中都有着广泛的应用，Yoonkyung Lee 等[8]应用多分类支持

向量方法对基因表达谱数据和卫星放射数据进行了分析. Dirong Chen 等[9]对多分类支持向量机算法的收敛性给出了理论分析. 窦智宙等[10]应用该方法对彩色癌症细胞的图像进行了分割. 康文雄等[11]应用小波分解和多分类支持向量机算法处理了脸谱识别问题.

2.1 多分类问题及贝叶斯决策

在多分类问题中,考虑如下样本$(\boldsymbol{x}_i, y_i)$, $i = 1,\cdots,m.$,其中, $\boldsymbol{x}_i \in \mathbf{R}^d$ 为输入样本,$y_i \in \{1,2,\cdots, k\}$ 为输出,代表了输入样本的分类情况. 主要目标是找到一种分类法则,使得分类法则 $\varphi(\boldsymbol{x})$ 可以尽可能精准的刻画输入 $\boldsymbol{x}_i$ 和输出 y_i 之间的关系. 假设样本为独立采样且满足分布 $P(\boldsymbol{x}, y)$, $p_j(\boldsymbol{x}) = P(Y = j \mid \boldsymbol{X} = \boldsymbol{x})$ 表示样本 $\boldsymbol{x}$ 被分到第 j 类的概率,当各类别之间的错分误差相等时,分类法则 $\varphi(\boldsymbol{x})$ 的损失可由如下函数定义:

$$l(y, \varphi(\boldsymbol{x})) = I(y \neq \varphi(\boldsymbol{x})).$$

其中:$I(\cdot)$ 为势函数,若自变量为真,则取值为 1,为假则取值为 0. 此时,最佳的分类法则为贝叶斯分类,即

$$\varphi_B(\boldsymbol{x}) = \arg \min_{j=1,\cdots,k} [1 - p_j(\boldsymbol{x})] = \arg \max_{j=1,\cdots,k} p_j(\boldsymbol{x}).$$

当错分误差不一样时,定义矩阵 $\boldsymbol{C}$,$\boldsymbol{C}_{jl}$ 表示第 j 被错分到 l 的惩罚. 显然有 $\boldsymbol{C}_{jj} = 0$, $j = 1,2,\cdots,k$, 此时

$$l(y, \varphi(\boldsymbol{x})) = \sum_{j=1}^{k} I(y = j) \left(\sum_{l=1}^{k} \boldsymbol{C}_{jl} I(\varphi(\boldsymbol{x}) = l) \right).$$

通过最小化风险函数得到的最优分类器为

$$\varphi_B(\boldsymbol{x}) = \arg \min_{j=1,\cdots,k} \sum_{l=1}^{k} \boldsymbol{C}_{lj} p_l(\boldsymbol{x}).$$

但在实际问题中,往往需要定义不同的错分惩罚,遇到这样的情况可以通过调整矩阵 $\boldsymbol{C}$ 中的元素值来表示不同的错分误差.

2.2 一对一(one against one)多分类支持向量机的数学模型

传统的 SVM 算法是专门处理小样本问题的,因此在考虑多分类问题时可以将问题转化成为多个二分类问题,这样的多分类支持向量机算法称为一对一的多分类算法[12].

一对一分类是在 k 类训练样本中构造所有可能的两类分类器,每个分类器仅仅在 k 类中的两类训练样本上训练,结果共构造 $k(k-1)/2$ 个分类器,测试样本经过各个分类进行分类,每次将样本判为某一类,最终得到判入次数最多的类为样本的最后结果. 但这种算法的运算量大,计算步骤很多,相应的运算时间也会长.

2.3 一对多(one against all)多分类支持向量机的数学模型

一对多方式的主要思想是构造 k 个二分类器,每个二分类器应用于所有样本,第 i 个二分类器将样本分为第 i 类和其他类,这 k 个二分类器组合起来就可以形成 N 分类的判决函数. 当一个样本数据输入时,依次用这 k 个二分类器进行判断,如果第 i 个二分类器的输出是属于第 i 类,而其他的分类器都输出为其他类,则判断该样本属于第 i 类.

多分类支持向量机算法最终转化为求解如下优化问题:

$$\min \sum_{i=1}^{n} \boldsymbol{L}(y_i)(\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{v}_i) + \frac{1}{2}\lambda \sum_{j=1}^{k} \| h_j \|_K^2.$$

其中:向量 $\boldsymbol{v}_i \in \mathbf{R}^k$ 记录分类结果,如果样本 $\boldsymbol{x}_i$ 属于第 j 类,则向量 $\boldsymbol{v}_i$ 第 j 个元素为 1,其他元素为 $-(k-1)^{-1}$;且映射 $\boldsymbol{L}(y_i)$ 第 j 个元素为 0,其他元素为 1 的向量,满足限定条件 $\sum_{j=1}^{k} \boldsymbol{f}_j(\boldsymbol{x}) = \boldsymbol{0}$. 该问题的最优解的算法求解过程本文借助 libsvm 工具包[13]中 SVMtrain 函数完成.

3 高校生源质量分析模型

基于多分类支持向量机的生源质量分析模型由数据采集、数据预处理、特征指标选取、数值归一化和学习过程组成,如图 2 所示.

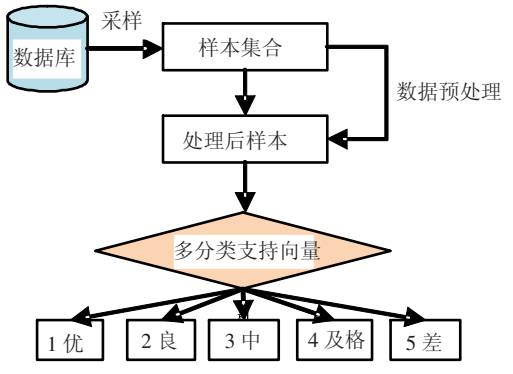

图 2 多分类支持向量机模型

1) 数据采集器. 按用户要求从数据库中采集数据;

2)数据预处理. 剔除残缺数据和奇异数据,如各省份特招生的样本. 由于各个省份的录取标准不同,分数之间的差距也很大,所以对高考成绩进行层次化处理,共分为 11 档次,如表 1 所示.

表 1　高考成绩分层处理

高考成绩区间	档次标号
[690, ∞)	1
[670, 690)	2
[650, 670)	3
[630, 650)	4
[610, 630)	5
[590, 610)	6
[570, 590)	7
[550, 570)	8
[530, 550)	9
[510, 530)	10
[0, 510)	11

3）特征指标选取. 从学生信息库中依照数据的属性和对问题的贡献选取了省份、高考成绩、所在省份排名、毕业时所在专业排名情况这 4 个指标组成了样本的有效信息（见表 2）.

表 2　有效样本

省区	成绩档次	省序	高考省内排名	总排名范围
四川	10	22	90 119	59.78%
山东	6	15	35 126	87.63%
湖北	10	17	33 132	13.96%
河北	8	3	15 578	47.75%
江苏	6	10	11 708	61.90%
湖南	7	18	9 236	48.60%
福建	8	13	2 179	21.05%
广西	5	20	2 178	67.86%

4）数值归一化处理. 由于各个指标之间数值的差别很大，所以对数值进行归一化处理，使得数值分布在[-1,1]之间：

$$\frac{2(data - \min(data))}{\max(data) - \min(data)}.$$

5）训练及学习过程. 对有标号的样本，根据其毕业时所在专业的排名进行分类（见表 3）：

表 3　带标号样本的分类情况

毕业时所在专业排名区间	所在等级
[0%, 10%]	1
(10%, 30%]	2
(30%, 60%]	3
(60%, 90%]	4
(90%, 100%]	5

4　测试实例与分析

数据来源主要是某大学某年级 3 000 余名本科学生的数据. 其中包括学生入学时的高考成绩以及在高中时期的其他信息（包括是否是三好学生、优秀学生干部，是否在学科竞赛中获奖等），另外一部分信息如生源所在省，录取时所在省份的排名情况以及高考成绩. 研究的数据量足够大，具有普遍性，使分析结果具有一定说服力.

4.1　多分类支持向量机模型的分类情况

实验基于高斯核函数及多项式核函数展开，并选取了不同核宽及不同的阶数进行了对比，在表 4 中，MSVM 表示多分类支持向量机模型，RBF_1 表示核宽为 1 的高斯核，POLY_1 表示阶数为 1 的多项式核. 在实验过程中，采用 10 倍交叉验证方法（10 - fold cross validation）测试各个模型的分类结果，即将数据集分成 10 组，轮流将其中 9 组做训练，1 组做测试，10 次的结果的均值作为对算法精度的估计，结果分别呈现了均值和波动值.

为比较多分类支持向量机的分类效果，采用多元统计中的判别分析作为对比试验. 判别分析是经典的多元统计方法，其主要思想是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法. 其基本原理是按照一定的判别准则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的待定系数，并计算判别指标. 据此即可确定某一样本属于哪一类. 实验结果见表 4.

表 4　实验结果（1）

分类模型	分类结果精度
MSVM（RBF_1）	0.774 72 ± 0.008 478 8
MSVM（RBF_5）	0.778 46 ± 0.012 719 4
MSVM（POLY_1）	0.753 74 ± 0.004 379 4
MSVM（POLY_3）	0.796 51 ± 0.037 178 1
Fisher linear discrimination	0.576 570 6
Fisher quadratic discrimination	0.700 674 3

通过实验可以看出，多分类支持向量机模型较传统判别分析模型有着更好的分类能力.

4.2　模型的改进

考虑对多分类支持向量机模型进行改进，在如上的实验中均假设对错误分类的惩罚是一致的，但在实际情况中并不是如此. 例如一种误差是将一名一等的学生错分为二等，即优秀的被判别成良好. 另一种误差是将一等学生错分到五等，即将优秀的学生判别成差下. 这两种错误的程度显然不一样，第二种错误较第一种错误要严重些，因此，可以调整惩罚矩阵 $\boldsymbol{C}_{jl}$，在本问题中最终用到的惩罚矩阵为

$$\boldsymbol{C}=\begin{pmatrix}0&1&2&3&4\\1&0&1&1&1\\1&1&0&1&1\\1&1&1&0&1\\1&1&1&1&0\end{pmatrix}\rightarrow\begin{pmatrix}0&1&2&3&4\\1&0&1&2&3\\2&1&0&1&2\\3&2&1&0&1\\4&3&2&1&0\end{pmatrix}.$$

含义为将第 i 类学生错分到第 j 的惩罚为 $|i-j|$.

基于此调整,训练得到新的分类器,实验结果如表 5 所示.

表 5 改进后的实验结果

分类模型	分类结果精度
MSVM(RBF_1)	0.871 253 ±0.007 583 5
MSVM(RBF_5)	0.882 312 ±0.000 365 8
MSVM(POLY_1)	0.860 205 ±0.004 539 1
MSVM(POLY_3)	0.887 214 ±0.004 581 7

由模型的实验结果可以看出,分类结果精度较调整前有了较大的提高.

5 结 论

1)本文基于多分类支持向量机算法建立了生源质量分析模型,该模型可以通过生源所在地、高考成绩,以及该成绩所在省的排名等信息,预测出该生源在大学毕业时学习情况的所处类别,因此可以对生源的质量优劣进行初步的分析和评价,对招生工作及相关政策的制定有一定的指导和帮助.

2)通过对惩罚矩阵进行合理的调整,模型的结果和可靠性都得到显著的提高.

参考文献:

[1] 姚金琢. 高等教育大众化与依法试行自主招生问题探析[J]. 中国高教研究, 2006 (11),90 - 91.

[2] 于春梅, 杨胜波, 陈馨,等. SVM 和基于 PCA、PLS 的 SVM 在非线性辨识中的比较研究[J]. 计算机应用研究,2004,24(6): 85 - 90.

[3] VAPNIK V. The nature of statistical learning theory [M]. [S.l.]: Springer Verlag, 1995.

[4] 韩媚,郭丹凤. SVM 的特征选择方法在高校就业预测中的应用研究[J]. 杭州师范大学学报:自然科学版,2009,8(5): 358 - 362.

[5] 向小东,宋芳. 基于核主成分与加权支持向量机的福建省城镇登记失业率预测[J]. 系统工程理论与实践,2009,29(1):73 - 80.

[6] 祁享年. 支持向量机及其应用研究综述[J]. 计算机工程,2004,30(10):6 - 9.

[7] SCHÖLKOPF, SMOLA A. Learning with Kernels[M]. Cambridge: MIT Press, 2002.

[8] LEE Y K, LIN Y, WAHBA G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data[J]. Journal of the American statistical Association, 2004,99 (465):67 - 381.

[9] CHEN D R, XIANG D H. The consistency of multicategory support vector machines[J]. Adv Comput Math, 2006,24(1 - 4): 155 - 169.

[10] 窦智宙, 平子良, 冯文兵,等. 多分类支持向量机分割彩色癌细胞图像[J]. 计算机工程与应用,2009,45 (20), 236 - 239.

[11] 康文雄,谢纪美,邓飞其,等. 基于小波分解和多分类支持向量机的脸谱识别[J]. 计算机测量与控制,2005,13(12):1390 - 1422.

[12] HSU C W, LIN C J. A Comparison of Methods for Multiclass Support Vector Machines[J]. IEEE Trans Neural Networks,2002,13(2):415 - 425.

[13] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[EB/OL]. [2008 - 07 - 25]. http://www.csie.ntu.edu.tw/ ~cjlin/libsvm.

(编辑 杨 波)