η-one-class 问题和 η-outlier 及其 LP 学习算法

陶 卿[1)] 齐红威[2)] 吴高巍[2)] 章 显[1)]

[1)](中国人民解放军炮兵学院 合肥 230031)

[2)](中国科学院自动化研究所 北京 100080)

摘 要 用 SVM 方法研究 one-class 和 outlier 问题. 在将 one-class 问题理解为一种函数估计问题的基础上,作者首次定义了 η-one-class 和 η-outlier 问题的泛化错误,进而定义了线性可分性和边缘,得到了求解 one-class 问题的最大边缘、软边缘和 ν-软边缘算法. 这些学习算法具有统计学习理论依据并可归结为求解线性规划问题. 算法的实现采用与 boosting 类似的思路. 实验结果表明该文的算法是有实际意义的.

关键词 one-class 问题;outlier;最大边缘;统计学习理论;支持向量机;线性规划问题;boosting

中图法分类号 TP18

η-One-Class Problems and η-Outliers with Their LP Learning Algorithms

TAO Qing[1)] QI Hong-Wei[2)] WU Gao-Wei[2)] ZHANG Xian[1)]

[1)](*New Star Research Institute of Applied Technology*, *Hefei* 230031)

[2)](*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100080)

Abstract In this paper, one-class and outlier problems are investigated by using the idea of Support Vector Machines. Based on regarding a one-class problem as the one to estimate a function, the generalization error for the one-class problem is defined for the first time. The linear separability, margin and optimal linear classifier are then defined and the regular SVM is reformulated into a framework for one-class problems. Each of the linear algorithms is motivated theoretically and they can be formulated as some linear programming problems. The proposed algorithms can be implemented by the techniques in boosting algorithms. Some synthetic and real experiments illustrate that the algorithms in this paper are practical and effective.

Keywords one-class problems; outliers; maximum margin; statistical learning theory; support vector machines; linear programming problems; boosting

1 引 言

目前,人们普遍认为统计学习理论[1,2]主要是研究从数据到分布的归纳机理问题,即问题的目标是分布意义下的最优性,而算法的目标是基于训练数据的. 如何设计机器学习算法从有限的样本集合得到分布意义下的最优,这是统计机器学习研究的主要内容.

对于 one-class 问题,我们一般作如下理解:已知一样本集合,构造一个能描述样本密度分布的二值模型. 如果某一样本处于较大密度的区域之内则认为它为正常样本,否则就认为它为 outlier. 一个自然的想法就是利用有限样本估计分布. 但在统计学

收稿日期:2003-04-19;修改稿收到日期:2004-03-10. 本课题得到国家自然科学基金项目(60175023)、安徽省自然科学基金项目(01042304)和安徽省优秀青年科技基金项目资助.**陶 卿**,男,1965 年生,博士,教授,研究领域为统计学习理论与算法、神经网络、应用数学. **齐红威**,男,1975 年生,博士,研究方向为机器学习和 Web 挖掘. **吴高巍**,男,1975 年生,博士,研究方向为神经网络、统计学习理论和支持向量机. **章 显**,男,1978 年生,硕士研究生,研究方向为支持向量机.

习理论中，一般不以估计分布为出发点，因为我们遵循一个原则[2]，即在解决一个给定问题时，要设法避免把解决一个更一般的问题作为其中间步骤.

one-class 问题是一种无监督的学习问题. 目前，基于统计学习理论的算法，如 SVM(Support Vector Machine)和 boosting[3]，在处理有监督的学习问题上卓有成效. 显然，能否设计统计学习算法解决 one-class 问题是人们非常关心的问题. 由于在 SVM 和 boosting 中，边缘(margin)的概念无论是在直观理解和算法设计，还是在理论分析中，地位和作用都非常重要. 因此，如何在 one-class 问题中提出这一概念成为研究中的关键问题.

1999 年，Tax 和 Duin 提出了一种 SVDD 方法(Support Vector Data Description)[4]求解 one-class 问题，其思路是寻求一个能把所有训练样本包围起来的最小超球，以此最小超球作为 one-class 问题的分类器. 在实际操作中，他们在覆盖的样本数和覆盖球的半径之间有所平衡，此时，位于覆盖球之外的训练样本则称为 outlier. 尽管文献[4]的研究符合人们的直观想象，但是从统计学习理论观点来看，文献[4]提出的仅仅是基于数据的学习算法，并没有从分布意义上去考虑 one-class 的泛化问题，因而也没有证据表明他们的算法是一种统计学习算法.

2001 年，Schölkopf 等人在文献[5]中提出将 one-class 问题看作是一种特殊的二分类问题. 有趣的是，他们的核算法与文献[4]的最小球覆盖算法相吻合. 但是他们的工作仍然存在着一定的不足，首先文献[5]没有给出 one-class 和 outlier 的确切定义，其次是对于线性可分情形，其算法用一个超平面作为 one-class 问题的分类器，这与人们的直观相距甚远.

在现实问题中，人们往往关心如下问题，即给定一些数据点和一个阈值，希望能找到一个函数，当样本点与此函数的误差小于给定的阈值范围时，认为它们是属于一类的. 但在实际中，用户有时无法给出有界区域的阈值，往往希望算法能自动地选取这个阈值. 本文尝试从理论层次解决这些问题. 值得提出的是，尽管 one-class 问题已引起了人们的充分注意[4~6]，但我们至今没有发现一个严谨的数学定义.

在本文中，我们将 one-class 和 outlier 问题理解为函数有界区域的估计问题，认为与一个待求函数的误差小于一个给定范围的点属于一类，而将误差大于这个范围的理解为 outlier，这个观点与文献[4～6]均不同. 我们希望这个待求函数满足在给定的阈值区域内尽量多地包括来源于独立同分布的数据点，据此我们定义了 η-one-class 的泛化错误. 这是首次明确地给出了 one-class 和 outlier 问题的定义，也是本文的主要创新，它为用统计学习理论进行分析奠定了必要的基础. 为了得到统计学习算法，我们沿用 SVM 从线性到非线性的思路，定义了边缘，讨论了泛化不等式和最大边缘算法，给出相应的优化问题和学习算法及其理论分析.

2　η-one-class 问题和 η-outlier

假设$\eta>0$是预先给定的允许精度，训练样本集为 $S=\{x_1,x_2,\cdots,x_l,x_i\in R^N,i=1,2,\cdots,l\}$，即 $x_1,x_2,\cdots,x_l$独立同分布于概率密度函数 $p(x)$. 记假设空间 H 为 R^N 上一些实值函数的集合. 对 $w\in R^N$，$p\in[1,+\infty]$，$\|w\|_p$表示 w 的 l_p范数.

定义 1(损失函数和泛化错误). 设 $f\in H$，定义

$$L(f(x),\eta)=\begin{cases}1, & \text{如果}|f(x)|\leq\eta\\ 0, & \text{其它}\end{cases},$$

称 L 为 η-one-class 问题的损失函数. 定义

$$R(f,\eta)=\int L(f(x),\eta)p(x)\mathrm{d}x$$

为 η-one-class 问题的泛化错误.

注意：在不考虑密度加权的情况下，这里的泛化错误可以粗略地解释为位于事先指定区域$|f(x)|\leq\eta$之外的服从分布 $p(x)$点的个数. 而使泛化性能最优则体现为使尽量多的服从分布 $p(x)$的点位于事先指定区域$|f(x)|\leq\eta$之中.

定义 2(经验风险). 设 $f\in H$，定义

$$R_{\mathrm{emp}}(f,\eta)=\frac{1}{l}\sum_{i=1}^{l}L(f(x_i),\eta)$$

为基于训练数据 $x_1,x_2,\cdots,x_l$的 η-one-class 问题的经验风险.

定义 3(线性可分 η-one-class 问题). 如果存在 $w\in R^N$和 $b\in R$，$\|w\|_p=1$，使得 $f(x)=w\cdot x+b$ 满足

$$|w\cdot x_i+b|\leq\eta,\quad i=1,2,\cdots,l,$$

则称 $x_1,x_2,\cdots,x_l$为线性可分的 η-one-class 问题，f 称为 η-one-class 问题的线性分类器.

定义 4(函数的 η-outlier). 设 $f\in H$，如果 x 满足 $\|f(x)\|>\eta$，则称 x 为分类器 f 的 η-outlier.

定义 5(η-one-class 问题的最优分类器与 η-outlier). 如果

$$R(f_0,\eta)=\min\left\{\int L(f(x),\eta)p(x)\mathrm{d}x,f\in H\right\},$$

则称 f_0 为 one-class 问题的最优分类器. f_0 在 $\{x_1, x_2, \cdots, x_l\}$ 中的 η-outlier 称为 η-one-class 问题的 outlier.

注意:在定义 4 中,函数的 η-outlier 与训练样本是没有关系的.从统计学习理论的观点来说,这里的 outlier 也应与训练样本无关.之所以如此定义是考虑到在实际问题中,往往需要从训练样本中发现 outlier.本文之后提到 outlier 时,都是指函数在训练集合中的 outlier.

定义 6(边缘). 设 $f(x)=w\cdot x+b$, $\|w\|_p=1$. 定义训练样本 x_i 关于 f 的 η 边缘为 $m(f,x_i,\eta)=\eta-|w\cdot x_i+b|$. 定义 f 的 η 边缘为 $m(f,S,\eta)=\min\{m(f,x_i,\eta), i=1,2,\cdots,l\}$.

边缘的几何意义见定理 1. 显然,$m(f,S,\eta)\geq 0$ 的充要条件是 f 为 η-one-class 问题 $\{x_1,x_2,\cdots,x_l\}$ 的线性分类器.

定义 7(最大边缘分类器). 令 $H=\{f: f(x)=w\cdot x+b, w\in R^N, b\in R\}$. 对线性可分问题,若 $f_0\in H$ 满足

$$m(f_0,S,\eta)=\max\{m(f,S,\eta), f\in H\},$$

则称 f_0 为 η-one-class 问题的最大边缘分类器.

本文的主要思路与 SVM 相一致.在理论分析方面,主要讨论 one-class 问题泛化错误的界,从而说明本文的算法可以在 PAC(Probably Approximately Correct)[8]框架下保证泛化性能最好.在对算法的描述和分析上,都是从线性可分问题着手.在实现上,本文主要考虑与参数有明确解释 ν-SVM 类似的 one-class 问题.与 SVM 不同的是,我们受 boosting 算法解决二分类问题的启发,对所得到的二次优化问题进行改进,使之成为线性优化问题.由于该优化问题形式上与 boosting 中的优化问题基本相同,从而可利用 boosting 算法中的技巧进行实现,本文采用 CG(Column Generation)方法进行实现.

3 学习算法

定理 1[9]. 设 $z\in R^N$,它不在超平面 $P=\{x: w\cdot x=0, w\in R^N\}$ 上,那么对任意 $p\in[1,+\infty]$,则

$$\frac{|\langle w,z\rangle|}{\|w\|_p}=\|z-P\|_q,$$

其中 $\|z-P\|_q$ 表示 z 到超平面 P 的在 l_q 范数意义下的距离,$\frac{1}{p}+\frac{1}{q}=1$.

显然,我们在第二节所定义的边缘是 l_q 范数意义下的距离.当 $p=2$ 时,它就是欧氏距离.所谓最大边缘分类器就是使得样本点离两个分界面 l_q 距离最大的分类器.

显然,对于线性可分问题,求解最大边缘分类器可归结为如下优化问题:

$$\begin{cases}\max\limits_{w,b,\gamma}\gamma\\ \|w\|_p=1,\ w\cdot x_i+b+\eta\geq\gamma,\\ \quad w\cdot x_i+b-\eta\leq-\gamma,\ i=1,2,\cdots,l\end{cases}\tag{1}$$

由于优化问题(1)中含有约束 $\|w\|_p=1$,它一般是非凸优化问题.与 boosting 在处理 one-class 问题的类似[6],我们令 $p=1$ 并进一步限制 $w\in R^N_+$,得到下面的线性优化问题

$$\begin{cases}\max\limits_{w,b,\gamma}\gamma\\ \|w\|_1=1,\ w\cdot x_i+b+\eta\geq\gamma,\\ \quad w\cdot x_i+b-\eta\leq-\gamma,\ i=1,2,\cdots,l\end{cases}\tag{2}$$

在优化问题(2)中,令 $\rho=\eta-\gamma$,优化问题(2)变为

$$\begin{cases}\min\limits_{w,b,\rho}\rho\\ \|w\|_p=1,\ w\cdot x_i+b\geq-\rho,\\ \quad w\cdot x_i+b\leq\rho,\ i=1,2,\cdots,l\end{cases}\tag{3}$$

众所周知,最小球覆盖算法实际上是求解下面的优化问题

$$\begin{cases}\min\limits_{x_0,R}R\\ \|x_i-x_0\|\leq R,\ i=1,2,\cdots,l\end{cases}$$

相比之下,优化问题(3)的意义非常明显,它实际上是在求能够覆盖所有样本点的一种最小带形区域.这种思想与最小球覆盖算法是完全类似的,只不过这里的区域是带形区域而不是球,区域中心是一条直线而不是一个点.另外,优化问题(3)还可以看成为一个特殊的回归问题,我们将在另文中对此进行详细讨论.

对于线性分类器情形的 one-class 问题,文献[6]的优化问题是

$$\begin{cases}\max\limits_{w,b,\rho}\rho\\ \|w\|_p=1,\ w\cdot x_i+b\geq\rho,\ i=1,2,\cdots,l\end{cases}\tag{4}$$

显然,优化问题(3)和(4)在形式上是有所不同的,造成这种不同的原因在于我们是将 one-class 问题理解为函数有界区域的估计问题,而文献[6]与文献[5]的观点是一致的,都认为 one-class 问题是训练点集与坐标原点的二分类问题.对于样本空间的 one-class 问题来说,显然优化问题(3)比(4)更直观和易于理解.另一方面,优化问题(3)和(4)在形式上又是基本相同的,只是多了一个不等式约束,这种相

似性启发我们可用与处理 boosting 中优化问题完全相同的方法来求解我们的优化问题.

设 C 是预先给定的大于 0 的常数，与 C-SVM 的思想一致，通过对区域大小和和区域包含样本点数目之间进行折中，可以得到下面的优化问题：

$$\begin{cases}\min\limits_{w,b,\rho,\xi,\xi^*} \rho+C\left(\sum\limits_{i=1}^{l}\xi_i+\sum\limits_{i=1}^{l}\xi_i^*\right)\\ w\cdot x_i+b\geq-\rho-\xi_i^*,\ w\cdot x_i+b\leq\rho+\xi_i,\\ \qquad i=1,2,\cdots,l\\ \|w\|_1=1,\ \xi_i\geq0,\ \xi_i^*\geq0,\ w\in R_+^N,\\ \qquad i=1,2,\cdots,l\end{cases}$$

由于 ν-SVM 中的参数意义比较明确，与 ν-SVM 类似，我们引进一个新的参数 ν，将上述优化问题变为

$$\begin{cases}\min\limits_{w,b,\rho,\xi,\xi^*} \nu\rho+\sum\limits_{i=1}^{l}\xi_i+\sum\limits_{i=1}^{l}\xi_i^*\\ w\cdot x_i+b\geq-\rho-\xi_i^*,\ w\cdot x_i+b\leq\rho+\xi_i,\\ \qquad i=1,2,\cdots,l\\ \|w\|_1=1,\ \xi_i\geq0,\ \xi_i^*\geq0,\ w\in R_+^N,\\ \qquad i=1,2,\cdots,l\end{cases}\tag{5}$$

参数 ν 的意义见后面定理 3. 优化问题(4)和优化问题(5)的不同之处在于用户预先设定的参数不同，一个是给定区域的具体范围，一个是给定了 outlier 个数. 正如文献[6]指出的那样，在非监督学习问题中参数的可解释显得十分重要. 可以看出，优化问题(4)和优化问题(5)正好可满足不同用户的需求.

令基本假设空间(或称之为弱分类器空间) $H_0=\{h_1,h_2,\cdots,h_N\}$，此时假设空间为 $H=\left\{\sum\limits_{i=1}^{N}w_ih_i(x):\ w_i\geq0,i=1,2,\cdots,N\right\}$. 记 $\Phi(x)=\sum\limits_{i=1}^{N}w_ih_i(x)$，我们可以得到下面求解非线性 one-class 分类器的优化问题：

$$\begin{cases}\min\limits_{w,\rho,\xi,\xi^*} \nu\rho+\sum\limits_{i=1}^{l}\xi_i+\sum\limits_{i=1}^{l}\xi_i^*\\ \Phi(x_i)\geq-\rho-\xi_i^*,\Phi(x_i)\leq\rho+\xi_i,\\ \qquad i=1,2,\cdots,l\\ \|w\|_1=1,\ \xi_i\geq0,\ \xi_i^*\geq0,\ w\in R_+^N,\\ \qquad i=1,2,\cdots,l\end{cases}\tag{6}$$

4　理论分析

为了对线性可分情形优化问题(1)进行理论解释，首先引进覆盖数的概念和相关定理[10].

定义 8.　设 F 是定义域为 X 的一些函数的集合，F 关于 $S=\{x_1,x_2,\cdots,x_l,x_i\in R^N,i=1,2,\cdots,l\}$ 的 γ 覆盖是一个由有限数目函数组成的集合 B，满足对任一 $f\in F$，都存在 $g\in B$，使得 $\max\limits_{1\leq i\leq l}|f(x_i)-g(x_i)|<\gamma$. 我们将数目最小的覆盖记为 $N(F,S,\gamma)$. 而 F 的覆盖数定义为 $N(F,l,\gamma)=\max\limits_{S\in X^l}N(F,S,\gamma)$.

尽管我们这里定义的边缘和泛化错误与 SVM 定义的有所不同，但与文献[10]中的定理 4.9 完全类似，我们可以得到以下定理.

定理 2.　设 F 是阈值型函数空间(对于分类问题，阈值型函数空间指的是假设空间的每一个函数由 F 中的函数阈值化构成)，$\gamma>0$. 设 $p(x)$ 是任一概率密度函数，$x_1,x_2,\cdots,x_l$ 来源于 $p(x)$ 的一组独立样本. 若 $f\in F,R_{\mathrm{emp}}(f,\eta)=0,m(f,S,\eta)\geq\gamma$. 那么当 $l>2/\varepsilon$ 时，下列不等式

$$R(f,\eta)\leq\frac{2}{l}\left[\log N(F,2l,\gamma/2)+\log_2\frac{2}{\delta}\right]$$

以概率 $1-\delta$ 成立.

定理 3.　令 $H=\{f:f(x)=w\cdot x+b,w\in R^N,\ \|w\|_1=1,b\in R\}$，则成立下列不等式

$$\log N(H,2l,\gamma/2)\leq1+\frac{144}{\gamma^2}(2+\ln l)\log\left(2\left[\frac{4}{\gamma}+2\right]l+1\right).$$

定理 3 是文献[13]中定理 2.3 当 $B=1$ 时的特殊情形，该定理的一般情形出自文献[14].

根据定理 2 和定理 3，类似于 SVM 的理论解释[10]和线性规划 boosting 方法的理论依据讨论[13]，本文的最大边缘分类器可在 PAC 意义下实现泛化性能界的优化. 另外，文献[10,12]通过定义辅助函数和辅助扩维函数空间，证明了软边缘算法实际上是一种特殊的最大边缘算法，这种证明方法很好地说明了软边缘算法理论依据. 完全与文献[13]类似，可以讨论软边缘优化问题(6)的统计学习理论依据问题. 至此，我们对本文所提出的一些算法进行了严谨的理论分析.

定理 4.　设优化问题(5)的解为 ρ_0,w_0,ξ_i 和 $\xi_i^*,i=1,2,\cdots,l$. 假设 $\rho_0\geq0$，则

$$\sum_{i=1}^{l}I(|f_{w_0}(x_i)|>\rho_0)\leq\nu\leq\sum_{i=1}^{l}I(|f_{w_0}(x_i)|\geq\rho_0),$$

其中 $I(\cdot)$ 是指示函数，当参数为真时取 1，而当参数为假时取 0.

证明.　令 $\nu_0=\sum\limits_{i=1}^{l}I(|w_0\cdot x_i+b)|>\rho_0)$. 显然，若 $|w_0\cdot x_i+b|>\rho_0$，必然有 ξ_i 或 ξ_i^* 非零. 根据

约束条件可知,ξ_i 和 ξ_i^* 中最多只能有一个不为 0. 因此我们可以假设前 ν_0 项的松弛变量 ξ_i 全不为 0,而其余的每一项 ξ_i 和所有的 ξ_i^* 均为零,此时显然有 $|w_0 \cdot x_i + b| \leq \rho_0, i = \nu_0 + 1, \nu_0 + 2, \cdots, l$. 由于训练集合只有有限个的元素,此时必然存在 ρ_1,满足 $\rho_0 < \rho_1 < \rho_0 + \min\{\xi_i, i = 1, 2, \cdots, \nu_0\}$.

一方面,显然有

$$w_0 \cdot x_i + b \geq -\rho_1, w_0 \cdot x_i + b \leq \rho_1, \quad i = \nu_0 + 1, \nu_0 + 2, \cdots, l.$$

另一方面

$$w_0 \cdot x_i + b \geq -\rho_1, w_0 \cdot x_i + b \leq \rho_1 + (\rho_0 - \rho_1 + \xi_i), i = 1, 2, \cdots, \nu_0.$$

注意到 $\rho_0 - \rho_1 + \xi_i > 0, i = 1, 2, \cdots, \nu_0$. 与优化问题(5)的约束条件相比较,可得

$$\nu\rho_0 + \left(\sum_{i=1}^{\nu_0} \xi_i\right) \leq \nu\rho_1 + \left(\sum_{i=1}^{\nu_0} (\rho_0 - \rho_1 + \xi_i)\right),$$

由此可得

$$\nu \geq \nu_0 = \sum_{i=1}^{l} I(|w_0 \cdot x_n + b)| > \rho_0).$$

同理 $\nu \leq \sum_{i=1}^{l} I(|w_0 \cdot x_n + b)| \geq \rho_0)$. 证毕.

定理 3 的结论和 ν-SVM 算法[7]的结论相似,也和处理 one-class 问题的 boosting 算法的结论相似[6],不过这里的优化问题与他们均不同. 根据定义 5,定理 3 我们看到,outlier 的个数不超过 ν,同时 ν 小于或等于 outlier 的个数和边界面上的点数之和. 正是由于参数具有这种解释,在算法实现中我们只讨论优化问题(6).

5 实 验

对本文线性优化问题的实现,我们采用了文献[13]中的 CG 方法. 而对弱分类器的设计问题,我们只是采用文献[6]中最简单的一种设计方法,即解如下的优化问题:

$$\max_{\alpha} \sum_{i=1}^{l} (d_i - d_i^*) h_\alpha(x_i), \ \|\alpha\|_1 = 1.$$

由于该优化问题具有如下形式的解

$$\alpha_q = \begin{cases} 1, & g_q = \max\limits_{q'} g_{q'} \\ 0, & \text{其它} \end{cases},$$

其中 $g_q = \sum_{i=1}^{l} d_i k_q(x_i)$,采用这种方法的好处在于所有弱分类器组成的集合此时就是 $\{k_1, k_2, \cdots, k_Q\}$.

由于本文的 one-class 分类器实际上是估计函数的有界区域,根据非线性情形下分类器 SVM 的表示形式,我们一般取核函数为 $k(x, x_n) = \exp\left(\frac{\|x - x_n\|^2}{\sigma^2}\right)$,而不是 SVM 中的 $k(x, x_n) = \exp\left(-\frac{\|x - x_n\|^2}{\sigma^2}\right)$. 此时理论分析与实验结果完全吻合,本节的前两个例子是人造的,用以直观说明算法的分类结果. 第三个例子是 one-class 问题用于股市异常波动个股的检测.

例 1. one-class 问题的线性分类器(图 1).

例 2. one-class 问题的非线性分类器,其中 $\sigma = 2, k(x, x_n) = \exp\left(\frac{\|x - x_n\|^2}{0.3 \cdot \sigma^2}\right)$(图 2).

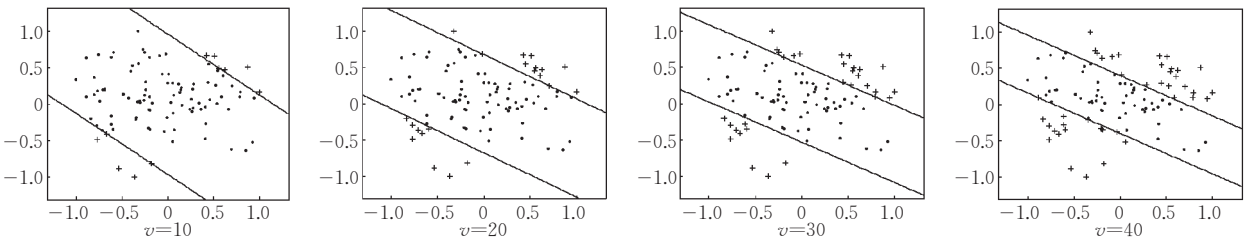

图 1 one-class 问题的线性分类器

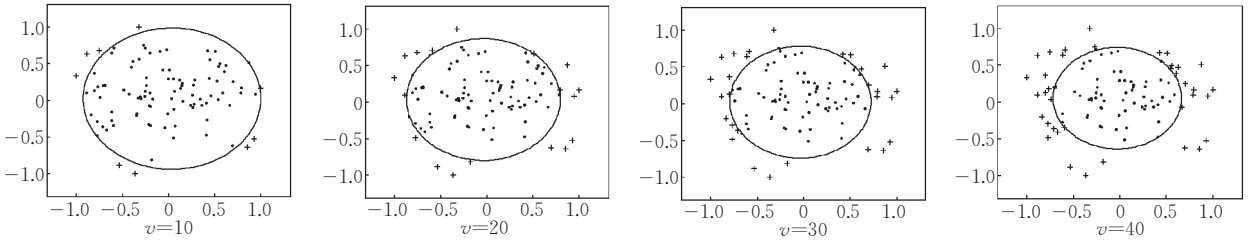

图 2 one-class 问题的非线性分类器

例 3.　股票的异常波动检测.

实验背景：波动是股票市场与生俱来的基本属性. 正是在价格的不断波动中，股票市场才发挥着基本的资源配置功能. 然而，股票价格的异常波动无论从宏观层面还是从微观层面来说都是对股票市场本身运作的巨大伤害[15,16]. 因此对股票市场的股价异常波动检测变得十分必要.

股票市场中的异常波动包括个股时间序列的异常波动和整个市场的横截面中的异常波动. 时间序列异常波动点，是指检测一个股票相对于自身的时间序列的异常波动点. 横截面异常波动点，是指在同一时间点上整个市场中的个别股票的价格异常波动，这是一个非时序问题，也是本实验要解决的问题.

异常波动源是指造成股价异常波动的影响因素，包括宏观因素、上市公司背景因素和庄家操纵或不能确认的其它因素等三种主要因素. 宏观因素一般指股票市场以外的国内外的相关的政治、经济、社会等造成股价异常波动的因素，如一个国家的经济形势、国家对宏观经济的调控政策、重大的社会事件(如战争)等. 上市公司背景因素指有关上市公司的生产、销售、资产运作等经营情况对本公司股价产生异常波动的因素，如公司的会计报表的公布、公司的股权结构的调整、资产并购行为等. 庄家操纵或不能确认的其它因素指除了宏观因素和公司背景因素之外造成股价异常波动的因素.

实验样本：在横截面中，存在着在价格出现异常波动时伴随着成交量的急剧放大的情况，因此在检测横截面中的异常点时，采用收益率(价格的衍生变量)和换手率作为指标最为自然，其中换手率是成交量和流通总股本的比值，指在一定时间内市场中股票转手买卖的频率，是反映股票流通性强弱的指标之一. 样本是深交所所有上市公司在某个横截面(通常取一个时间段)上的收益率和换手率. 本实验的数据来自 http://finance.yahoo.com/?u，上市公司公告来自中国证券监督管理委员会网址：http://www.csrc.gov.cn，人民网：http://www.people.com.cn 和数码证券网：http://www.my0578.com.

基本观点：当某支股票的收益率和(或)换手率严重偏离其它股票的这些指标时，这支股票就是异常的. 当使用 one class 分类器分类时，outlier 就是异常的股票.

实验结果：我们选择所有深交所上市公司 2000.6.1～2000.6.30 横截面的数据. 令 $v=10$ 和 $\sigma=2$，$k(x,x_n)=\exp\left(\frac{\|x-x_n\|^2}{0.3\cdot\sigma^2}\right)$，用我们的方法检测出 outliers，再与当时的公告进行比较，查出造成异常的异常波动源，结果见表 1.

表 1　横截面异常波动表(20000601～20000630)

异动	股票名称	上涨下跌	个股信息或宏观因素	公布时间	类型
1	000033	下跌	新都酒店股权变动警示性公告：本公司第一大股东中国东方信托投资公司拟将所持本公司法人股 6825 万股中的 6615 万股法人股转让深圳市卢堡工贸有限公司，其有关转让合同已签署.	20000526	B
2	000035	下跌	中科健重大事项公告：2000 年 6 月 7 日，中华人民共和国对外经济贸易合作部发布了关于暂停进口原产于韩国的手持无线电话机和聚乙烯的公告. 由于本公司生产的科健移动电话主要零部件仍依赖韩国进口，因此，从即日起本公司的移动电话原材料暂时无法解决，对本公司的生产经营将产生重大影响. 提请投资者注意投资风险.	20000609	B
3	000039	上涨			C
4	000554	上涨			C
5	000806	下跌	银河科技年度利润分配公告：以总股本 16275.60 万股为基数，向全体股东每 10 股送 3 股派现金 0.75 元(扣税 10 送 3)，股权登记日：2000 年 6 月 6 日，除权除息日：2000 年 6 月 7 日，红股上市交易日：2000 年 6 月 8 日.	20000601	B
6	000817	上涨	辽河油田公告：鉴于本公司 A 股股票交易已连续三天达到涨幅限制，根据有关规定，董事会特此提示：本公司目前生产经营正常；到目前为止，本公司无应披露而未披露的信息；请广大投资者注意投资风险.	20000607	C
7	000866	上涨	扬子石化派息公告：以总股本 233000 万股为基数，向全体股东每 10 股派现金 0.80 元(扣税后 10 派 0.64 元).	20000621	B
8	000895	下跌	双汇发展公积金转增股本公告：以总股本 22490 万股为基数，向全体股东公积金每 10 股转增 3 股，股权登记日：2000 年 6 月 5 日，除权日：2000 年 6 月 6 日，转增股可流通部分起始交易日：2000 年 6 月 7 日.	20000530	B
9	000908	下跌	天一科技重大事项公告：公司于 2000 年 5 月 26 日与国防科技大学计算机学院签署合作协议，公司主要提供资金和后勤保障、市场开拓等，国防科大计算机学院负责技术工作和技术支持.	20000527	B

注. 异动类型指包括宏观因素、个股消息因素和其他因素等三种异常波动源，其中“A”指宏观因素，“B”指个股消息因素，“C”指非 A 和非 B 的因素.

实验结论：从表 1 中可以看出：(1)从异常波动源的角度看，用 one-class 分类器检测出的 9 次异常波动中，7 次(除第 3 和第 4 次外)有公告作为直接证据. (2)许多股票发生消息泄漏事件，即股价往往

在公司公告前就发生异动,如异常波动 2 和 7.综上,我们认为 one-class 分类器可为异常波动的检测和分析提供参考作用.

对于本实验,作者声明:本实验的结果只用于算法测试,除此之外没有任何其它意图.

6 结 论

与关于 one-class 问题的其他文献相比,本文的主要贡献在于首次从统计学上明确定义了 one-class 问题的泛化错误,并提出了相应的统计学习算法进行求解.实际上,我们建立了与 SVM 完全类似的 one-class 问题统计学习算法体系.

致 谢 非常感谢审稿人的有益建议!

参 考 文 献

1 Vapnik V.. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

2 Vapnik V.. Statistical Learning Theory. Addison-Wiley, 1998

3 G. Rätsch. Robust boosting via convex optimization[Ph D dissertation]. University of Potsdam, 2001

4 Tax D., Duin R.. Data domain description using support vectors. In: Proceedings of the European Symposium on Artificial Neural Networks, 1999, 251～256

5 Schölkopf B., Platt J., Shawe-Taylor J., Smola A. J., Williamson R. C.. Estimating the support of a high-dimensional distribution. Neural Computation, 2002, 13(7): 1443～1471

6 Rätsch G., Mika S., Schölkopf B., Müller K. R.. Constructing boosting algorithms from SVMs: An application to one-class classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 9(4): 1184～1199

7 Schölkopf B., Smola A., Williamson R. C., Bartlett P. L.. New support vector algorithms. Neural Computation, 2000, 12:1207～1245

8 Valiant L. G.. A theory of the learnable. Communications of the ACM, 1984, 27(11): 1134～1142

9 Mangasarian O. L.. Arbitrary-norm separating plane. Operation Research Letters, 1999, 24(1):15～23

10 Cristianini N., Schawe-Taylor J.. An Introduction to Support Vector Machines. Cambridge: Cambridge University Press, 2000

11 Schawe-Taylor J., Bartletett P. L., Williamson R. C., Anthony M.. Structure risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory, 1998, 48(10): 1926～1940

12 Schawe-Taylor J., Cristianini N.. On the generalization of soft margin algorithms. IEEE Transactions on Information Theory, 2002, 48(10): 2721～2735

13 Demiriz A., Bennett K., Shawe-Taylor J.. Linear programming boosting via column generation. Machine Learning, 2002, 46(1): 225～254

14 Zhang T.. Analysis of regularized linear functions for classification problems. IBM: Technical Report RC-21572, 1999

15 Pan Deng. Research on investor's behavior in Chinese stock market[Ph. D. dissertation]. Shanghai Jiaotong University, Shanghai, 2000(in Chinese)
(攀 登. 中国证券投资者行为研究[博士学位论文]. 上海交通大学,上海, 2000)

16 Wu Wen-Feng. Trading Volume-Evolved Stock Price and Its Dynamic Analysis[Ph. D. dissertation]. Shanghai Jiaotong University, Shanghai, 2001(in Chinese)
(吴文锋. 基于成交量进程的股价动力学分析[博士学位论文]. 上海交通大学,上海,2001)

TAO Qing, born in 1965, Ph. D., professor. His research interests are in the area of statistical learning theory and algorithms, neural networks and applied mathematics.

QI Hong-Wei, born in 1975, Ph. D.. His research interests include Machine Learning and Text/Web Mining.

WU Gao-Wei, born in 1975 Ph. D.. Currently he is a post-doctor fellow in Institute of Computing Technology, Chinese Academy of Sciences, P. R. China. His research interests focus on neural networks, statistical learning theory and SVM.

Zhang Xian, born in 1978. Currently he is a master degree candidate in New Star Research Inst of Applied Tech, Hefei, P. R. China. His research interests focus on SVM.

Background

Our lab is focused on research and application of statistical leaning theory and algorithms. The theme is to develop and apply theories and methods of statistical machine learning for pattern recognition and data mining. Our current research includes the statistical learning theoretical motivation for learning algorithms, support vector machines and the fast geometric implementation of the learning algorithms. Our research is partially supported by the National Science Foundation of China《Geometrical theory on SVM and implementations》. This paper tries to employ the geometric idea in SVM to solve the simplest unsupervised learning problem——one-class problems. It illustrates that the geometrical theory on SVM for supervised learning problems can be adapted to obtain statistical learning algorithms for unsupervised problems.