孤立点一类支持向量机算法研究

田 江 顾 宏
(大连理工大学电子与信息工程学院 大连 116023)

摘 要：一类支持向量机将数据样本映射到高维空间，通过与坐标原点保持最大间隔的特征超平面检测孤立点。实际应用中算法对坐标原点的选择依赖性较强，检测性能受数据样本的分布影响较大；将算法转化为求解二类问题在一定程度上克服了这些不足，但其带来的数据不平衡问题受到现实中孤立点样本稀少或者不存在的影响。该文提出了“孤立点一类支持向量机”算法，并在此基础上设计了一种无监督的孤立点检测方法。分别基于超平面距离和概率输出大小定义两种孤立点异常程度，设定不同权值合并两种异常程度输出，将获得的可疑孤立点特征信息引入算法；在特征空间划分距离可疑孤立点最大间隔的超平面，分析在全部样本上的预测输出大小进而交互更新两部分的数据样本。在 UCI 数据集上进行了仿真实验，数据结果表明了该文方法能有效的提高检测率，降低误报率；同时样本交叉更新提高了检测的稳定性。

关键词：孤立点挖掘；一类支持向量机；癌症检测
中图分类号：TP181 **文献标识码**： A **文章编号**：1009-5896(2010)06-1284-05
DOI: 10.3724/SP.J.1146.2009.00861

Outlier One Class Support Vector Machines

Tian Jiang Gu Hong
(*School of Electronic and Information Engineering*, *Dalian University of Technology*, *Dalian* 116023, *China*)

Abstract: One-Class Support Vector Machines (OCSVMs) distinguish outliers by computing a hyper-plane in feature space. The choice of the origin as separation point is arbitrary, which affects the decision boundary, and the distribution of samples has impact on the performance. Expanding the algorithm into solving two-class classification problems overcomes the drawbacks to a certain degree. However, the class imbalance problem is serious and the labeled outliers are rare or even non-existing. In this paper, a new “Outlier OCSVM” is proposed and a framework is designed for unsupervised outlier detection. Respectively scored by distance from hyper-plane and probabilistic output value, two definitions of outlier degree are presented. After picking out some suspicious outliers via combining the two criterions of outlier degree, the adjusted “Outlier OCSVM” starts the training operations, two parts of the dataset are updated interactively through comparison of the outputs. Experiment results on benchmark datasets show that the method can effectively improve the detection rate and reduce false positive rate, easy and reliable.

Key words: Outlier mining; OCSVMs; Cancer detection

1 引言

孤立点检测用于发现不具备一般数据特性的数据样本，进而发现潜在的有用信息。孤立点的产生可能是因为度量或系统执行错误所导致的，通常在数据库中所占比例远低于正常数据。许多数据挖掘算法试图使孤立点的影响最小化，或者排除它们；但是由于一个人的噪声可能是另一个人的信号，这可能导致重要的隐藏信息的丢失，因为孤立点本身可能非常重要。孤立点检测可以应用到很多领域，如信用卡欺诈检测、安全系统、医学诊断、网络入侵检测和信息检索等，是数据挖掘领域的一个重要研究方向[1−4]。

支持向量机的成功促使其被应用到孤立点检测中，其中一类支持向量机得到了研究人员的重视[5,6]。与传统支持向量机解决二类分类问题相比，孤立点检测本质上是一类分类问题。Scholkopf 等[7,8]提出一类支持向量机算法，在特征空间计算与坐标原点保持最大间隔的分类超平面，将孤立点与其他正常数据样本区分出来。一类支持向量机的一个不足在于其性能依赖于原点的选择[9]，另外所有的孤立点都被

2009-06-09 收到，2009-10-16 改回
国家自然科学基金(60605022)资助课题
通信作者：田江 tianjiang@gmail.com

假设位于和正常样本不同的区域，或者说孤立点和正常样本具有定性的区别。在复杂的应用问题中，某些孤立点可能同正常样本具有相近的分布，导致这样的孤立点无法被正确识别。文献[10]在一类支持向量机的启发下，将一类问题转化为二类问题，实验结果显示在特定数据集中具有更强的稳定性，在一定程度上克服了噪声的影响；但其整体的分类性能没有提高。此外，这些研究成果对一类支持向量机的改进均是假定已经获得了信息足够丰富的孤立点标签样本，能够反映出孤立点整体的分布情况，如果条件不满足则会影响实际的检测效果。在许多实际孤立点检测应用中，由于孤立点样本获取的成本很高或者难度很大，研究人员只能得到很少的孤立点样本或者只有正常数据样本；另外孤立点可能出现在数据分布的任何区域内，而转化为二类分类问题会带来新的数据不平衡问题。这些不利因素都会影响到算法的检测效果，限制了算法的推广应用。

为克服上述缺点和不足，本文在一类支持向量机的基础上设计了一种旨在提高其泛化能力的无监督孤立点检测方法，尝试在无监督框架下引入可疑孤立点信息来解决问题。提出了孤立点一类支持向量机算法，在计算过程中引入“可疑孤立点”信息，特征空间内训练数据与可疑孤立点保持最大间隔，从而有针对性地提高分类超平面的精确性。设计了探测“可疑孤立点”的算法，缓解了实际中孤立点信息匮乏的问题。提出了一种样本交叉算法以避免正常样本被错判为孤立点，充分利用异常程度较高的孤立点样本以提高算法的稳定性。在 UCI 数据集上进行了实验，在相同的参数条件下与一类支持向量机进行了比较；结果表明本文方法经过交叉更新后检测结果逐渐达到稳定，在提高检测率的前提下同时降低了误报率。

2　理论基础

支持向量机将数据样本映射到一个高维空间，在这个空间建立最大间隔分类超平面以分离不同类别的数据。Scholkopf[7,8]等提出了一类支持向量机算法，通过划分超平面将正常样本与坐标原点保持最大间隔，坐标原点本质上是唯一的第二类数据样本。算法的主要目的在于估计高维空间中大量样本出现的区域，决策函数能够正确预测大多数样本所在的区域。

在孤立点检测的其他方法中，需要定义孤立点的异常程度，并以此来进行检测。这里分析一类支持向量机对孤立点预测输出的异常程度大小，设 $f(x)$ 表示决策函数输出大小，做如下假设和定义：

如果 $f(x)>0$，则 x 是正常样本；如果 $f(x)<0$，则 x 为孤立点。

定义 1　存在两个样本 x 和 x'，决策函数输出均小于 0，即 $f(x)<0$ 和 $(x')<0$ 同时成立。如果 $|f(x)|>|f(x')|$，那么说 x 较 x' 的异常程度更高，或者说“可疑”程度更高。

3　孤立点一类支持向量机

3.1　探测可疑孤立点

本文提出的孤立点一类支持向量机利用到部分孤立点信息，需要从训练样本中探测出一些可疑孤立点，即按照定义 1 给出的可疑程度较高的样本点。为提高系统的稳定性，选取了两种方式来刻画样本和分类超平面的远离程度，分别是与超平面的距离和概率决策输出的大小。在一类支持向量机的优化过程中可以计算样本点和特征超平面的距离，使用文献[11]的方法能够将决策输出数值转化为概率形式。两种孤立点异常程度的计算方法均将标签型的预测输出转化为小数形式，反应了样本属于正常点还是孤立点的某种程度。

基于单一的模型，通常很难判断计算获得的可疑孤立点是否会严重依赖于某一个特定的训练模型。为此需要选取多个不同复杂度的模型进行训练，进而通过合并输出计算可疑孤立点。对一类支持向量机来说，通过设定不同的折中参数 ν 可以获得具有不同复杂度的模型。针对某一个特定样本是否属于可疑孤立点，重点考虑两个因素：被检测到的频率和平均可疑程度。被检测到的频率越高，这个样本就越有可能是孤立点。同样，可疑程度越高表明这个样本属于孤立点的可能性越高。通过设置权值可以平衡两种因素的可疑程度，再结合检测到的频率可以得出最可疑的孤立点，具体如算法 1。

算法 1　探测可疑孤立点

输入：训练样本。

输出：可疑孤立点的中心点信息。

步骤 1　设定折中参数，读取数据样本，设训练样本个数为 n 。

步骤 2　设 $k=1$，这里 $(1\le k\le 10)$，表示取 10 组不同复杂度的训练模型。

步骤 3　计算当前模型的复杂度参数为 $(0.5+0.1k)\nu$，其取值范围为 0.5ν ~1.5ν 。

步骤 4　训练模型，得到对应的决策函数。

步骤 5　分别按照超平面距离和概率输出标准计算 $n\nu$ 个可疑孤立点。

步骤 6　按照定义 1 对预测输出从大到小排序，并且按照下式等比例设定权值：

$w=1+p/(n\nu)$，其中 $p=1,2,\cdots,n\nu$,是可疑孤立点排序后的标号。

对中间过程异常程度高的样本位置和权值大小进行记录。

步骤 7 $k=k+1$，返回到步骤 3。

步骤 8 合并训练模型的预测结果，对相同样本的输出进行相加，对结果从小到大排序。

步骤 9 输出排序后的前 $n\nu$ 个可疑孤立点对应的中心点信息。

3.2 孤立点一类支持向量机

实际应用中孤立点样本较少或者不存在，而在无监督检测算法中利用孤立点信息可以改善一类支持向量机的不足，文献[12]将数据样本与部分孤立点样本的中心点保持最大间隔，降低对原点的依赖，但其存在的问题是需要提前获取部分已知的孤立点信息。本文算法通过计算得到一些“可疑”程度最高的孤立点，利用这些点的信息计算一个特征点，进而构造超平面使其他数据样本与特征点保持最大间隔。检测原理如图 1 所示，算法在优化过程中能引入部分异常程度较高的数据样本信息，进而调整分类超平面，尽可能地分隔孤立点与正常点。

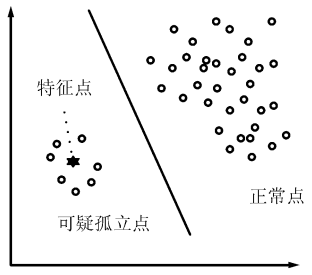


图 1 孤立点一类支持向量机

对于孤立点一类支持向量机，在具体计算中，原始的优化问题如下：

$$\left.\begin{aligned}&\min \quad \frac{1}{2}\|w\|^2+\frac{1}{\nu l}\sum_i\xi_i-\rho\\&\text{s.t.} \quad \langle w\cdot(x_i-m_0)\rangle\ge\rho-\xi_i,\xi_i\ge 0\end{aligned}\right\} \tag{1}$$

其中 m_0 为由部分异常程度较高的数据样本合成的特征点，ν 是训练误差和模型复杂度的折中参数。为求解优化问题，引入拉格朗日因子 $\alpha_i,\beta_i>0$，构造拉格朗日方程：

$$\begin{aligned}L(w,\xi,\rho)=&\frac{1}{2}\|w\|^2+\frac{1}{\nu l}\sum_i\xi_i-\rho\\&-\sum_i\alpha_i(w\cdot(x_i-m_0)-\rho+\xi_i)-\sum_i\beta_i\xi_i\end{aligned} \tag{2}$$

分别将式(2)对 w,ξ,ρ 求偏导，令其偏导数为 0，可得

$$\begin{aligned}&w=\sum_i\alpha_i\cdot(x_i-m_0),\\&\alpha_i=1/(\nu l)-\beta_i\le 1/(\nu l),\ \sum_i\alpha_i=1\end{aligned} \tag{3}$$

将式(3)代入式(2)，通过消元可得

$$\begin{aligned}L(w,\xi,\rho)=&\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j(\langle x_i\cdot x_j\rangle+\langle m_0\cdot m_0\rangle\\&-\langle x_i\cdot m_0\rangle-\langle x_j\cdot m_0\rangle)\end{aligned} \tag{4}$$

引入核函数，进而将式(1)转化为对偶问题。

$$\left.\begin{aligned}\min\ &\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j(k(x_i,x_j)+k(m_0,m_0)-k(x_i,m_0)\\&-k(x_j,m_0))\\\text{s.t.}\ \ &0\le\alpha_i\le 1/(\nu l),\ i=1,\cdots,l\\&\sum_{i=1}^n\alpha_i=1\end{aligned}\right\} \tag{5}$$

通过计算可得决策函数如下：

$$f(x)=\operatorname{sgn}\left(\sum_i\alpha_i k((x_i-m_0),(x-m_0))-\rho\right) \tag{6}$$

上述孤立点一类支持向量机在优化过程中引入可疑孤立点信息，有助于提高孤立点检测的性能。另外传统一类支持向量机优化过程中使用全部数据样本，这样孤立点样本很可能会成为支持向量，从而造成决策超平面向孤立点倾斜进而增大孤立点的漏检率。新算法将可疑孤立点与正常样本分开，降低了可疑孤立点样本在计算中被当作支持向量的可能性，从而在整体上能够提高检测性能。

3.3 动态交叉孤立点检测框架

使用“孤立点一类支持向量机”设计了一种充分利用可疑孤立点信息并根据预测输出对训练数据样本动态更新的检测框架方法，具体流程如图 2 所示。将数据样本分为两份，分别用“普查样本”和“待定样本”表示；它们在检测框架方法中的作用不同，但在选取时是没有特别要求的，均为随机选择。在可疑孤立点的探测过程中，由于孤立点的数据样本分布不同，会存在一定的鉴别误差，而伪孤立点会严重影响检测性能。为了增强稳定性，这里设计了实验数据样本交叉更新的过程，在两组数据中将利用预测输出的异常程度最大的样本点进行交叉替换更新，然后进入新一轮的迭代运算。具体描述如算法 2。

算法 2 数据交叉更新

输入：全体数据样本。

输出：部分样本更新后的数据集。

步骤 1 读取数据样本，设定孤立点一类支持向量机的参数和交叉比例。

步骤 2 输入中心点数据样本。

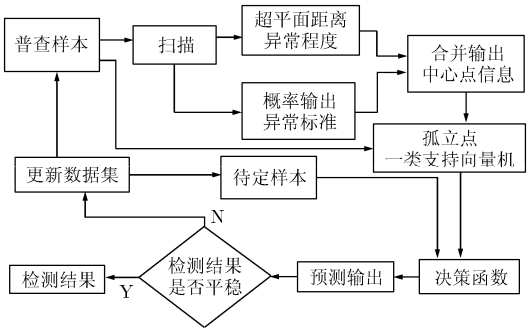


图 2 孤立点检测框架

步骤 3　在普查样本上用孤立点一类支持向量机训练。

步骤 4　分别对普查样本和待定样本进行预测，得到输出数值。

步骤 5　按定义 1 在两组样本上计算异常程度最高的样本点，进行交叉更新。

步骤 6　对更新后的数据进行下一次迭代计算。

4　实验分析

为评估算法的有效性，在 UCI 数据集上进行了实验。使用检测率和误报率来度量孤立点检测算法的性能，在后面的实验分析中，用 TPR 表示检测率，用 FPR 表示误报率。本文主要在高斯核函数下进行了算法的仿真实验，通过选择标准 c=hd 选取了一组参数[13]，其中 $h\in\{0.1,0.2,0.5,0.8,1.0\}$，$d$ 是数据维数。本文实验用到的软件基于 Libsvm 开发。

4.1 乳癌数据集(wisconsin breast cancer dataset)

该数据集有 666 条记录，其中良性的有 458 条记录，恶性的有 241 条记录，由 9 个数值属性刻画。选取 40(8%)条恶性记录和 444(92%)良性记录构造不平衡数据集，具体执行过程中等分成普查样本和待定样本，各为代表孤立点数据的 20 条恶性记录，和 222 条作为正常样本的良性记录。参数设置为 $\nu=0.1$，按选择标准选取 5 个核函数参数。首先分析孤立点检测算法中迭代操作对检测结果的影响，将实验迭代运算 10 次，实验结果分别在图 3 给出。

从图中可以看出，在不同的参数下本文方法表现出不同的性能，在前 3 组参数下实验结果比较接近；然而后两组参数下变化幅度较大，经过波动后达到比较理想的效果。后两组参数下的性能波动较大，这是因为核宽度参数影响到了分类超平面的变化，预测出了一些伪孤立点；但是经过样本交叉更新后的算法，能保证系统在不同参数下均达到较高的检测性能。实际中算法在迭代过程中能够充分利用可疑程度最高的数据样本，进一步保证了算法的稳定性。此外，可以看出在第 5 次迭代运算后，孤立点检测的 TPR 和 FPR 逐渐趋于平稳，检测算法到达稳定状态，经过多次交叉后数据集中的可疑孤立点基本确定。

实际算法中根据前后几次检测结果的均值和方差判断是否达到稳定，这里取到达稳定状态后的平均检测结果与一类支持向量机算法进行比较分析，实验结果比较如图 4 所示。实验结果表明，本文方法在相同的参数下，孤立点检测率高于一类支持向量机，而相应的误报率则低于一类支持向量机。在 5 组实验参数下，检测率有大幅提高，由一类支持向量机的 64%提高到 83.5%；而平均误报率由一类支持向量机的 4.86%减低到 2.52%。在单独的每组参数模型下，本文算法的性能均高于一类支持向量机，表现出算法较高的稳定性。

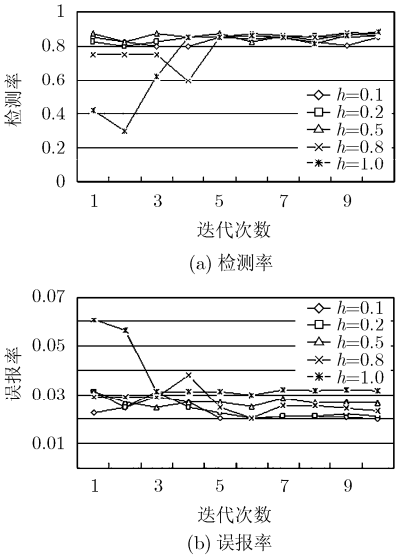
4.2 淋巴系造影术数据集(lymphography dataset)

这个数据集包含 148 条记录，每条记录包含 19 个分类属性，记录分为 4 类；类 1 和类 4 共占整个数据集的 4.05%，可以看成孤立点样本。将数据集中的正常样本和孤立点样本等分成 2 份，构成本实验的数据集。参数设置为 $\nu=0.05$，核函数参数选择同上个实验一致，具体检测率和误报率的比较结果如图 5 所示，其中本文方法的检测结果取稳定后的平均数值。

从图 5 中可以看出，本文方法在 5 组参数实验的结果中，在 TPR 上有 1 组与一类支持向量机保持一致，另外 4 组实验均得到了提高；对于 FPR 本文方法也表现出了较好的性能，有 4 组数据低于一类支持向量机，另外 1 组数据的性能稍差。与一类支持向量机的结果比较，本文方法的平均 TPR 从 53.3%提高到 83.3%；而与此相对应的平均 FPR 则从 25.6%降低到 17.6%。本文方法的优势主要体现在能够有效降低整体数据的FPR，进而提高稳定性。

5　结论

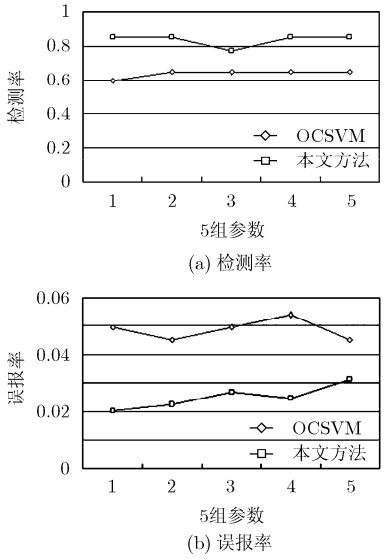
本文提出了孤立点一类支持向量机算法，在特征空间有针对性地刻画与可疑孤立点保持最大间隔的分类超平面；设计了权衡考虑分别基于超平面距离和概率输出两种异常程度的可疑孤立点刻画标准，并在此基础上提出了一种根据数据异常程度动态更新数据样本的检测框架方法。在真实数据集上的实验证明了方法的有效性，实验结果表明本文方法具有更好的孤立点检测性能，提高了检测率，同时降低了误报率。算法应用简单，对两组数据样本分布没有特殊要求，均为随机选取。另外在检测中考虑多次可疑孤立点的影响，数据集动态交叉更新算法增强了孤立点检测的稳定性。



(a) 检测率

(b) 误报率

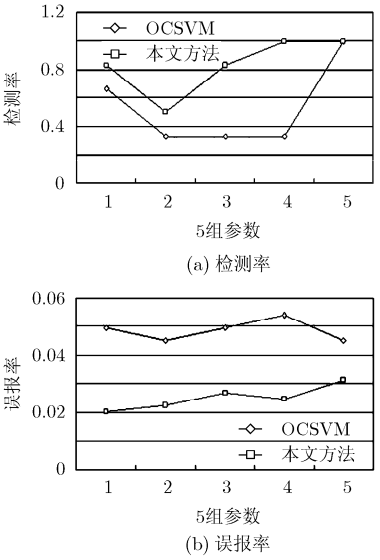
图 3 交叉更新的检测率和误报率



(a) 检测率

(b) 误报率

图 4 检测率和误报率比较



(a) 检测率

(b) 误报率

图 5 检测率和误报率比较

参 考 文 献

[1] Han J and Kamber M. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2006: 451–458.

[2] 倪巍伟, 陈耿, 陆介平, 吴英杰, 孙志挥. 基于局部信息熵的加权子空间孤立点检测算法[J].计算机研究与发展, 2008, 45(7): 1189–1194.
Ni W W, Chen G, Lu J P, Wu Y J, and Sun Z H. Local entropy based weighted subspace outlier mining algorithm[J]. *Journal of Computer Research and Development*, 2008, 45(7): 1189–1194.

[3] 薛安荣, 鞠时光, 何伟华, 陈伟鹤. 局部孤立点挖掘算法研究[J]. 计算机学报, 2007, 30(8): 1455–1463.
Xue A R, Ju S G, He W H, and Chen W H. Study on algorithms for local outlier detection[J]. *Chinese Journal of Computers*, 2007, 30(8): 1455–1463.

[4] 庞彦伟, 刘政凯. 一种自动抑制孤立点的子空间学习方法[J]. 电子与信息学报, 2008, 30(1): 176–179.
Pang Y W and Liu Z K. Automatically outlier-resisting subspace learning[J]. *Journal of Electronics and Information Technology*, 2008, 30(1): 176–179.

[5] Giacinto G, Perdisci R, Del Rio M, and Roli F. Intrusion detection in computer networks by a modular ensemble of one-class classifiers[J]. *Information Fusion*, 2008, 9(1): 69–82.

[6] Chandola V, Banerjee A, and Kumar V. Anomaly detection: A survey[J]. *ACM Computing Survey*, 2009, 41(3): 1–58.

[7] Scholkopf B, Williamson R C, Smola A J, Shawe-Taylor J, and Platt J. Support vector method for novelty detection[J]. *Advances in Neural Information Processing Systems*, 2000, 12(3): 582–588.

[8] Scholkopf B, Platt J C, Shawe-Taylor J, Smola A J, and Williamson R C. Estimating the support of a high-dimensional distribution[J]. *Neural Computation*, 2001, 13(7): 1443–1471.

[9] Eskin E, Arnold A, Prerau M, Portnoy L, and Stolfo S. Applications of Data Mining in Computer Security[M]. Norwell Massachusetts:Kluwer Academic Publishers, 2002: 77–87.

[10] Manevitz L M and Yousef M. One-class SVMs for document classification[J]. *Journal of Machine Learning Research*, 2002, 2(2): 139–154.

[11] He J R, Li M J, Li Z W, Zhang H J, Tong H H, and Zhang C S. Pseudo relevance feedback based on iterative probabilistic one-class SVMs in web image retrieval, In: Advances in Multimedia Information Processing - Pcm 2004, Pt 2, Proceedings[C]. Lecture Notes in Computer Science, 2004, 3332: 213–220.

[12] Scholkopf B, Platt J, and Smola A J. Kernel method for percentile feature extraction[R]. Microsoft Research Ltd, 2000.

[13] Munoz A and Moguerza J M. Estimation of high-density regions using one-class neighbor machines[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(3): 476–480.

田 江： 男，1979 年生，博士生，研究方向为机器学习、数据挖掘等.

顾 宏： 男，1961 年生，教授，博士生导师，从事自动控制理论及应用、机器学习等研究.