基于 ICA 与 SVM 的孤立点挖掘模型*)

彭红毅[1] 蒋春福[2] 朱思铭[2]

(华南农业大学理学院 广州 510642)[1] (中山大学数学与计算科学学院 广州 510275)[2]

摘 要 本文提出一种基于独立成分分析(ICA)与支持向量机(SVM)的孤立点挖掘模型——ISOM 模型,用 ICA 对观测到的多维随机向量进行独立成分分解,用 SVM 估计独立成分的密度函数,克服了传统孤立点挖掘方法的一些缺点,为数据挖掘提供了一种有效的方法,并通过实验验证了该模型的合理性与正确性。

关键词 孤立点,ICA ,SVM,密度函数估计

Outlier Mining Model Based on ICA & SVM

PENG Hong-Yi[1] JIANG Chun-Fu[2] ZHU Si-Ming[2]

(College of Science, South China Agricultural University, Guangzhou 510642)[1]

(Department of Mathematics, Sun Yat-sen University, Guangzhou 510275)[2]

Abstract ISOM, Outlier Mining Model Based on ICA & SVM, is presented in this paper. This model transforms an observed multidimensional random vector into mutually independent components by ICA and estimates independent components' density function by SVM. Overcoming the defects of traditional outlier mining, the model of ISOM provides an efficient method for data mining, and its correctness and reasonableness ars also validated by the experiment results in this paper.

Keywords Outlier, ICA(independent component analysis), SVM(Support Vector Machine), Estimation of density function

1 引言

孤立点是数据集中的小部分对象,这些对象与数据中的一般行为或数据模型有着明显的不同。从实际来看,它能用于欺诈监测,例如探测不寻常的信用卡使用或电信服务。此外,在市场分析中可用于确定极低或极高收入的客户消费行为,或在医疗分析中用于发现多种治疗方式的不寻常的反应。这样,探测和分析孤立点、挖掘隐藏信息成为一项重要的数据挖掘任务,被称为孤立点挖掘。早期的孤立点探测多见于统计领域,近来,研究人员又提出了各种各样的方法,大致可以分为基于距离的方法、基于深度的方法、基于密度的方法、基于聚类的方法和基于神经网络的方法[1~6]。这些方法实际上是假定数据之间是互相独立的,但实际生活中很多数据存在一定的相关性,这样获取的孤立点在某种程度上不合理。Methmed K.[7]介绍了一种挖掘孤立点的方法,此方法事先假定数据服从高斯分布,实际上有很多数据并不服从高斯分布,因此这个缺点极大地限制了它的应用。P. J. De Groot[8]用主成分分析法研究了孤立点的挖掘问题,但主成分的方法只是基于二阶统计特性,仅仅去掉了数据间的相关性,却不能保证处理后的数据互相独立。

ICA 是一种基于高阶统计和信息理论的新的统计方法,旨在将观察到的数据变量分解成尽可能线性独立的成分,是一种多用途的统计方法。ICA 最早由 Jutten C 和 Herault J[9]提出,后来 Yogesh S[10]提出了一种简化的 ICA 方法,Andras K 和 Janos C[11]对 Fast ICA 进行了相关研究和应用并给出了代码。ICA 的兴起为有效挖掘孤立点提供了基础。根据观察到向量的维数 k 和独立成分向量的维数 m 的不同情形,ICA 可以分为三类:标准 ICA 问题($k=m$),不完备 ICA 问题($k>m$)和过完备 ICA 问题($k<m$)。本文讨论 $k\geq m$ 的情形。

SVM 由 Cortes 和 Vapnik 于 1995 年首先提出[12],是近年来机器学习的一项重大成果。SVM 基于结构风险最小化原理,与传统神经网络相比,支持向量机不仅结构简单,而且各种技术性能尤其是泛化能力明显提高,是求解模式识别和密度函数估计问题的一种有效方法[13,14]。

本文主要研究数据之间存在相关关系时孤立点的挖掘方法。文章第 2 节提出了基于 ICA 与 SVM 的孤立点挖掘模型(简称 ISOM 模型),对指标筛选算法进行了设计,并介绍了一种用 SVM 估计密度函数的方法;第 3 节介绍和分析了实验结果;最后是小结。

2 ISOM 模型及算法设计

2.1 ISOM 模型

基于 ICA 与 SVM 的孤立点挖掘模型 ISOM 如图 1 所示。

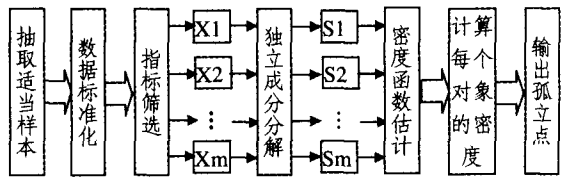

图 1 ISOM 模型

*)本文得到国家自然科学基金项目(10371135)资助。**彭红毅** 博士生,研究方向:数据挖掘、人工智能。**蒋春福** 博士生,研究方向:金融统计。**朱思铭** 教授,博士生导师,研究方向:人工智能与计算机网络、动力系统、混沌理论。

在孤立点挖掘模型中，用户一般需先确定要进行孤立点挖掘的 k 个数据指标，并按适当比例从数据集中随机抽取样本，接着对样本数据进行标准化，然后从这 k 个指标中筛选出 m 个线性无关的指标，再进行独立成分分解，之后估计独立成分的密度函数，最后，根据各独立成分的密度值就可以确定孤立点。本文采用 SVM 中样条密度估计的方法估计各独立成分的密度，进而得出各独立成分的联合密度，并按照其大小找出孤立点。

2.2 指标筛选

设样本容量为 n，记标准化处理后的指标为 $X_1, X_2, \cdots, X_k$，这里 $X_i=(x_{i1}, x_{i2}, \cdots, x_{in})^T, i=1,2,\cdots,k$，再从这 k 个指标中筛选出 m 个线性无关的指标。记 $X=(X_1, X_2, \cdots, X_k)^T$，$C=XX^T$，一般地，样本容量 n 要远远大于指标数量 k，因此矩阵 C 所需的存储空间远比矩阵 X 要小，并且我们还注意到矩阵 C 与矩阵 X 的线性无关向量之间存在着一定的对应关系，即若由从矩阵 C 的第 $i_1, i_2, \cdots, i_m$ 列中抽取第 $i_1, i_2, \cdots, i_m$ 行组成的列向量线性无关，亦即若矩阵 C 的第 $i_1, i_2, \cdots, i_m$ 列和第 $i_1, i_2, \cdots, i_m$ 行组成的子矩阵的行列式非零，则必有 $X_{i1}, X_{i2}, \cdots, X_{im}$ 也线性无关，下面给出具体的指标筛选算法：

①初始化 $G=C, j=k, d[i]=0, i=1,\cdots,k$；

②如果 $\det(G)\geqslant\varepsilon_G$，其中 $\det(\cdot)$ 表示矩阵行列式值，ε_G 是一个很小的正数，则转第⑤步；

③如果 $\deg(G)<\varepsilon_G$，则 $d[j]=1$，将 G 重新置为 G 去掉第 j 行、j 列后的矩阵，$j=j-1$；

④如果 $j\geqslant 1$，则转第②步，否则转第⑤步；

⑤$h=\sum_{i=1}^{k} d[i]$，如果 $d[j]=0, j=1,\cdots,k$，则 X_j 是我们要找的线性无关的行向量，一共有 $(k-h)$ 个。

2.3 密度函数的支持向量机估计

不妨设由上述指标筛选算法筛选后的 m 个线性无关的指标为 $X_1, X_2, \cdots, X_m$，并假定它们由 m 个非高斯独立成分线性混合而成。记 $X^{(m)}=(X_1, X_2, \cdots, X_m)^T$，利用 Fast ICA 算法[11,15]将其样本数据进行分解，可得 $X^{(m)}=AS$，其中 A 是 m 阶可逆矩阵，$S=(S_1, S_2, \cdots, S_m)^T$，$S_i (i=1,2,\cdots,m)$ 是第 i 个独立成分。下面给出单个独立成分的密度样条逼近方法：

Step1 将样本映射[0,1]到区间：

假设我们已知独立成分 s 的样本数据为，$s_1, s_2, \cdots, s_n$，令 $s_{\min}=\min_{1\leqslant i\leqslant n}\{s_i\}$，$s_{\max}=\max_{1\leqslant i\leqslant n}\{s_i\}$，$\tilde{s}=s_{\min}-(s_{\max}-s_{\min})/n$，作变换 $t=(s-\tilde{s})/(s_{\max}-\tilde{s})$，显然 t 是 s 的单调递增函数，根据 s 的样本数据我们可以得到变换后独立成分 t 的样本数据 $t_1, t_2, \cdots, t_n$。

Step2 构造若干三元组：构造经验分布函数

$$F_n(t)=\frac{1}{n}\sum_{i=1}^{n}\theta(t-t_i),$$

其中 $\theta(u)=\begin{cases}1, u\geqslant 0\\ 0, u<0\end{cases}$

定义 $\varepsilon_i=\sqrt{\frac{1}{n}(F_n(t_i)+\delta)(1-F_n(t_i)+\delta}$，其中 δ 是一个很小的正数，则我们就可构造若干三元组 $(t_1, F_n(t_1), \varepsilon_1), \cdots, (t_n, F_n(t_n), \varepsilon_n)$。

Step3 在像空间中定义相应的回归问题：

利用 Step2 中构造的三元组数据来逼近函数 $F(t)=\sum_{i=1}^{n}\beta_i K(t_i,t)+b$，$t_i$ 为训练集中的一个向量，其中 β_i 为某一个实数，$K(t_i,t)$ 为一个核函数。训练集中对应于非零 β_i 值的向量为支持向量。将 β_i 改写成下列形式：$\beta_i=\alpha_i^*-\alpha_i$，其中 $\alpha_i^*\geqslant 0, \alpha_i\geqslant 0$。

Step4 在像空间中构造核函数：

$$K(t_i,t)=|t_i-t|^2\frac{(t_i\wedge t)^3}{12}+|t_i-t|\frac{(t_i\wedge t)^4}{8}+\frac{(t_i\wedge t)^5}{20}+\frac{t_i^2t^2}{4}+t_it,$$

其中 $(t_i\wedge t)$ 表示两个值 t_i 和 t 中的小者。

Step5 计算交叉核函数：

$$K(t_i,t)=\frac{t_i^2t}{2}+x_i+\frac{t_i^2t(t_i\wedge t)}{2}-(2t_it+t_i^2)\frac{(t_i\wedge t)^2}{4}+(2t_i+t)\frac{(t_i\wedge t)^3}{6}-\frac{(t_i\wedge t)^4}{8}$$

Step6 求解支持向量和对应的系数：

在满足约束

$$\begin{cases}|y_i-\sum_{j=1}^{n}(\alpha_j^*-\alpha_j)K(t_i,t_j)-b|\leqslant\varepsilon-\xi_i\\ \alpha_i^*\geqslant 0, \alpha_i\geqslant 0, i=1,2,\cdots,n\end{cases}$$

的条件下最小化泛函

$$W(\alpha,\xi_i)=\sum_{i=1}^{n}\alpha_i+\sum_{i=1}^{n}\alpha_i^*+c\sum_{i=1}^{n}\xi_i,$$

其中 $\varepsilon=\max_{1\leqslant i\leqslant n}\{\varepsilon_i\}+\delta_1$，$\delta_1$ 是一个很小的非负数，C 是一个给定值，$\xi_i=\varepsilon-\varepsilon_i, i=1,\cdots,n$。

Step7 变换后独立成分密度函数的支持向量逼近：

由 $\beta_i=\alpha_i^*-\alpha_i, i=1,\cdots,n$ 得其中 N 个不为零的系数，设为 $\beta_k^0=\alpha_i^*-\alpha_i$ 和相应的支持向量 $t_k, k=1,\cdots,N$，从而得变换后成分的密度函数的支持向量逼近：

$$q(t)=\sum_{k=1}^{N}\beta_k^0K(t_k,t)$$

Step8 独立成分的密度函数估计：

$$p(s)=q\left(\frac{s-\tilde{s}}{s_{\max}-\tilde{s}}\right)/(s_{\max}-\tilde{s})$$

根据上述算法可以近似估计出每个独立成分的密度函数 $p(S_i), i=1,\cdots,m$，从而得到独立成分的联合密度函数 $p(S)=\prod_{i=1}^{m}p(S_i)$。对数据集上的每个对象，如果有缺失值，则需要先进行缺失值的估计，然后通过 $S=A^{-1}X^{(m)}$ 求出独立成分，进而可得到它的独立成分的密度值，其中缺失值的估计方法可参见文[15]。

最后，按密度值进行升序排列，密度值最小的 M 个数据（也可用数据总量乘以适当比例）对象即为孤立点。

3 实验结果

下面实验利用上海 2005 年一季度最新财务指标 172 条完整数据（截止至 4 月 22 日，下载网址：http://www.stock2000.com.cn）验证本文提出的模型。一共选取股票的 4 个财务指标，分别为：每股收益摊薄（元）X_1，调整后的每股净资产（元）X_2，每股净资产（元）X_3，净资产收益率（%）X_4，实验中采用 SAS9.0 进行编码，并在 PIV1.7G，256M 的 PC 上运行。

表 1 是当选择的孤立点个数为 8 时，孤立点挖掘结果（表中 name 表示股票名称，density 表示股票 4 个指标进行独立成分分解后各独立成分的联合密度值）。表 2 是 4 个财务指标的平均值与标准差列表。表 3 为 4 个指标相关系数列表。图 2 为 172 个样本进行独立成分分解后，按独立成分联合密

度值升序排列后的散点图。

表 1 孤立点挖掘结果列表

name	X_1	X_2	X_3	X_4	density
ST 博讯	−0.015	0.08	0.08	−19	9.169E-7
千金药业	0.19	11.18	11.45	1.67	5.652E-4
伊泰 B 股	0.38	3.41	3.67	10.3	1.147E-3
济南钢铁	0.39	4.05	4.05	9.58	1.393E-3
恒源煤电	0.38	6.78	6.78	5.6	2.122E-3
马应龙	0.29	10.59	10.61	2.71	2.269E-3
烟台万华	0.18	2.22	2.22	10.82	4.260E-3
云天化	0.3136	4.36	4.4	7.13	4.757E-3

表 2 四个财务指标的平均值与标准差

	X_1	X_2	X_3	X_4
平均值	0.08008	3.22628	3.29674	2.30820
标准差	0.08589	1.60271	1.58628	2.93456

表 3 各指标相关系数

	X_1	X_2	X_3	X_4
X_1	1.00000	0.46550	0.46405	0.72010
X_2	0.46550	1.00000	0.99493	0.10048
X_3	0.46405	0.99493	1.00000	0.09949
X_4	0.72010	0.10048	0.09949	1.00000

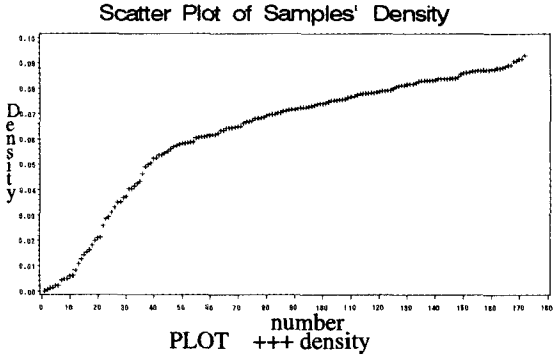

图 2 样本独立成分联合密度的散点分布图

从表 3 可以看出，X_2 与 X_3 具有很大的相关性，根据指标筛选方法，我们最后选择指标 X_1、X_3 和 X_4 进行独立成分分析。从表 1、表 2 中可以看出，第一个孤立点“ST 博讯”指标 X_4 明显偏离中心值，第二个孤立点“千金药业”的指标 X_1，X_4 明显偏离中心值，但偏离程度不如第一个大，找出的前面 8 个孤立点至少都有一个指标明显偏离中心值，与实际情况基本吻合。从图 2 我们也可以看出，样本大部分集中在密度值比较大的地方。需要说明的是，孤立点挖掘的结果，只是给用户提供一个参考，只有用户才能最后确定真正的孤立点。

结论 挖掘孤立点，发现有价值的隐藏信息，是数据挖掘的一项重要内容。传统的孤立点挖掘方法实际上是假定数据之间是互相独立的，或事先假定数据服从高斯分布，但实际生活中很多高维数据之间存在一定的相关性，且很多数据不服从高斯分布，这些缺点限制了传统孤立点挖掘方法的应用。本文提出的 ISOM 模型，先用 ICA 对数据进行独立成分分解，再用 SVM 估计各独立成分的密度函数，克服了传统孤立点挖掘方法的局限性，为数据挖掘提供了一种有效的方法。实验结果也验证了本文提出的 ISOM 模型的有效性与合理性。

参 考 文 献

1 Liu Xiao-Hui. Strategies for outlier analysis. Birkbeck College University of London, 2000
2 Edwin M K, Raymond T Ng. Algorithm for Mining Distance-Based Outliers in Large Databases. In: Proc. of the 24th VLDB Conf. New York, USA, 1998
3 Johanna H, Rocke D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 2004, 44: 625～638
4 Ester M, et al. A Density-Based Algorithm for Discovering Clusters in large spatial databases. In: Proc. of 2nd Intl. Conf. on Knowledge Discovery and Data Mining, 1996
5 Bayarri M J, Morales J. Bayesian measures of surprise for outlier detection. Journal of Statistical Planning and Inference 2003, 111: 3～22
6 Bullen R J, et al. Outlier detection in scatterometer data: neural network approaches. Neural Networks (in press)
7 Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms. Tsinghua University Press, 2003
8 De Groot P J, Postma G J, et al. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. Analytica Chimica Acta, 2001, 446: 71～83
9 Jutten C, Herault J. Independent component analysis versus PCA. In: Proc. of European Signal Processing Conf. 1988. 287～314
10 Yogesh S. A simplified approach to independent compoent analysis. Neural Comput & Applic. 2003, 12: 173～177
11 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces. LNCS, 2001, 2234: 271～281
12 Cotes C, Vapnik V. Support Vector networks. Machine Leaning, 1995, 20: 273～295
13 Bartlett P L, Shawe-Taylor J. Generalization performance on support vector machines and other pattern classifiers. In: B. Sholkopf, C. Burges, A. Smola, eds. Advances in Kernel Methods-Support Vector Learning, Cambridge, MA: MIT Press, 1999
14 Vapnik V N. Statistical Learning Theory. Publishing House of Electronics Industry, 2004
15 彭红毅，朱思铭，蒋春福. 数据挖掘中基于 ICA 的缺失数据值的估计. 计算机科学，2005，32(12)：203～205

(上接第 163 页)

总结 本文重点研究了基于混沌神经网络和混沌映射的混沌伪随机序列的性能。用理论与计算机仿真实验相结合的方法对混沌序列的随机性、平衡性、相关性和线性复杂度等特性进行了系统的分析，并给出了相应的特性曲线。分析结果表明，基于混沌神经网络和混沌映射的混沌伪随机序列具有十分理想的随机特性和相关特性。由于混沌序列的产生非常方便，数量众多，因此可以用来替代 m-序列，以满足 CDMA 通信对大容量的需求。混沌序列的线性复杂度高，不容易破译，因此除了用于扩频通信之外，还可以作为传统密码学中的密钥来使用。

参 考 文 献

1 肖国镇著. 伪随机序列及其应用. 北京：国防工业出版社
2 Diffie W, Hellman M E. New directions in cryptography. IEEE Trans. on Information Theory, 1976, 22 (6): 644～654
3 van Schyndel R G, Tirkel A Z, Svalbe I D. Key independent watermark detection. IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 1999. 580～585
4 Eggers J J, Su Jonathan K, Girod B. Public key watermarking by eigenvectors of linear transforms. In: European Signal Processing Conference, Tampere, Finland, 2000. 428～435
5 Chua L O, Roska T. The CNN paradigm. IEEE Trans. CAS-I, 1993, 40: 47～156
6 Civalleri P P, Gilli M. On dynamic behaviour of two-cell cellular neural networks, Int. J. Circ. Th. Appl., 1993, 21: 451～471
7 丘水生，陈艳峰，吴敏，等. 混沌加密的若干问题与新的加密系统方案. 见：2002 中国非线性电路与系统学术会议论文集. 中国：深圳，2002，11：174～179
8 王育民. 混沌密码序列使用化问题. 西安电子科技大学学报，1997，24(4)：560～562
9 Tohur K, Akio T. Pseudonoise sequence by chaotic nonlinear and their correlation properties. IEICE Trans commun, 1993, E97-B (8): 855～862
10 Rueppel R A. Linear complexity and random sequences, Advances in Cryptology. EURO CRYPT'85 (LNCS 219), 1986. 167～188
11 何振亚，张毅锋，卢宏涛. 细胞神经网络动态特性及其在保密. 通信中的应用. 通信学报，1999，20(3)：59～67