基于 ICA 与 MViSOM 的孤立点挖掘模型*)

彭红毅[1] 蒋春福[2] 朱思铭[3]

(华南农业大学理学院 广州 510642)[1] (深圳大学数学与计算科学学院 深圳 518060)[2]

(中山大学数学与计算机科学学院 广州 510275)[3]

摘 要 本文提出了一种基于独立成分分析(ICA)与改进的可视化诱导自组织映射(MViSOM)的孤立点挖掘模型——IMVOM模型,该模型用 ICA 方法对观测到的多维随机向量进行独立成分分解,得到一个独立成分数据集,然后用改进的 MViSOM 方法取得数据的可视化。该模型充分结合"人类擅长于模式识别的能力"与"电脑擅长于大量地记忆、快速地计算的能力"的双方优点进行孤立点的挖掘,避免了对高维数据内部结构的复杂探测,从而克服了高维数据集孤立点挖掘过程中的一些困难。实验结果也验证了所提模型的合理性。

关键词 孤立点,ICA,MViSOM

Outlier Mining Model Based on ICA and MViSOM

PENG Hong-Yi[1] JIANG Chun-Fu[2] ZHU Si-Ming[3]

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

(College of Mathematics and Computing Science, Shenzhen University, Shenzhen 518060)[2]

(College of Mathematics and Computing Science, Sun Yat-sen University, Guangzhou 510275)[3]

Abstract IMVOM, Outlier Mining Model Based on ICA & MViSOM, is presented in this paper. This model firstly transforms an observed multidimensional random vector into mutually independent components by ICA, and then achieves visibility of high-dimensional data by MViSOM. Combined the pattern recognition capacity of human being with the calculating capacity of computer, this model can finish mining the outliers by avoiding of detecting the complex inner structure of data and overcoming some difficulties of outlier mining of high-dimensional data. In the end, the proposed model's correctness and reasonableness are also validated by the experiment results in this paper.

Keywords Outlier, ICA(independent component analysis), MViSOM(Modified Visualization-Induced Self-Organizing Maps)

1 引言

孤立点挖掘是数据挖掘研究中的一项重要内容,其目标是发现数据集中行为异常的数据对象。在许多数据挖掘应用中,有时异常信息比常规模式更有价值。

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[1~3]。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,这样获取的孤立点具有某种程度上的不合理。M. Kanatardzic[4]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,而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[5]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,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。同时,通常高维数据的内部结构很难探测,这也极大地限制了传统孤立点挖掘方法的应用。论文的主要工作就是克服传统孤立点挖掘方法的这两个弱点。

ICA 是一种基于高阶统计和信息理论的新的统计方法,目的是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 C. Jutten 和 J. Herault[6]提出,K. Andras 和 C. Janos[7]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一定的基础。

自组织映射(Self-organizing maps, SOM)是一种竞争的无指导学习方法,可以将任意的高维数据映射到一维或二维的网络图,为数据挖掘提供了非常有用的高维数据可视化技术[7,8]。H. Yin 在 SOM 的基础上提出了一种可视化诱导自组织映射(Visualization-Induced SOM, ViSOM),它使用与 SOM 同样的网络结构[9]。与 SOM 相比较,ViSOM 对群聚数

*)本文得到国家自然科学基金项目(10371135)资助。彭红毅 博士,研究方向:数据挖掘、人工智能。蒋春福 博士,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。

据具有更好的高维数据可视化分类功能[9~12]。彭红毅等[13]在 ViSOM 的基础上进行了进一步的改进，提出了一种改进的 MViSOM 方法，MViSOM 方法能保持数据之间一定的拓扑结构不变性，同时取得较好的数据可视化效果。

本文第 2 节在 ICA 与 MViSOM 的基础上提出了一个孤立点挖掘模型，称之为 IMVOM 模型，该模型能有效地去除数据之间的相关性，并能避免高维数据内部结构的复杂探测进行孤立点的挖掘；第 3 节是实验结果；最后是小结。

2 IMVOM 模型

基于 ICA 与 MViSOM 的孤立点挖掘模型 IMVOM 的流程如图 1 所示。

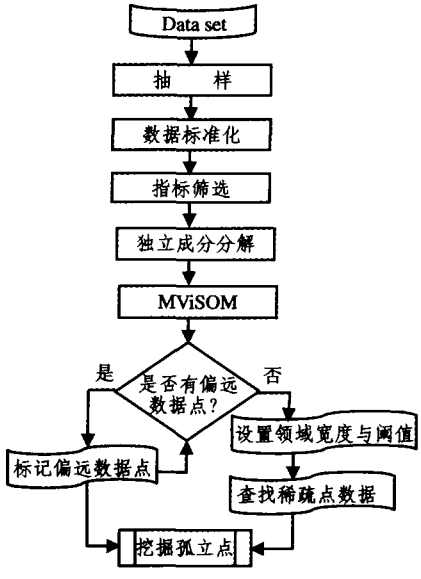

图 1　IMVOM 孤立点挖掘模型的流程图

模型 IMVOM 的算法实现可按以下步骤进行：

步骤 1　从数据集中随机抽取适量样本；

步骤 2　数据标准化处理；

步骤 3　指标筛选，详细的指标筛选算法可参见文[14]；

步骤 4　独立成分分解；

步骤 5　训练 ViSOM 网络；

步骤 6　用 MViSOM 将数据映射到二维空间图形详细的 MViSOM 方法可以参见文[13]；

步骤 7　标记所有偏远的数据点；

步骤 8　对于非偏远的数据点，设置一个邻域宽度与阈值，判断其是否为稀疏点数据，并对判断为稀疏点的数据进行标记；

步骤 9　挖掘出所有的孤立点，即找出标记为偏远的数据与标记为稀疏点的数据。

对于一个利用模型 MViSOM 已经得到二维可视化图形的数据集，偏远的数据点就是图形上远离其它数据点的数据，我们可以对此类数据点作好标记，并认为其为孤立点数据。

设数据集样本数为 T，训练的网络为 ViSOM $M\times N$ 网络，数据 $x(t)$ 的映射点为 (i,j)，对于 IMVOM 算法中的步骤 8，我们设置如下的算法：

①计算每个网格的平均数据点个数 $k=T/((M-1)\times(N-1))$；

②根据实际设置邻域宽度 σ 与阈值 $f=4\sigma^2\cdot c\cdot k$（$c$ 是一个正常数，例如取 $c=0.25$），计算以 (i,j) 为中心，邻域宽度为 σ 的正方形区域数据点个数 num；

③如果 $num<f$，则数据 $x(t)$ 为稀疏点数据。

实际上，偏远数据点与稀疏数据点的区别是相当模糊的，这是因为孤立点本身并没有一个明确界限的定义。算法中的邻域宽度 σ 与常数 c 并没有很明确的规定，用户可以根据实际需要自行设定。

3 实验结果

实验中从某航空公司某年度会员乘机行为数据表中整理出会员年度乘机数据表，为节省运算开销，只给出会员乘机次数大于 10 的记录共 1082 条，数据表中包含四个字段：会员编号 num(取值从 1 到 1082)，会员年度乘机次数 X1，会员年度飞行总里程 X2(单位：公里)，会员年度总票价 X3(单位：元)。每条记录包含对三个指标进行孤立点挖掘。为了保证实验的可行性和有效性，对 IMViSOM 方法中的一些参数作了如下规定：

(1)$M=30$，$N=30$，二维神经元矩阵为 30×30；

(2)排序阶段：学习的最大次数为 2000；学习率的初始值为 0.9；学习率的最终值为 0.05；邻域的初始值为 20；邻域的最终值为 3；

(3)收敛阶段：学习的最大次数为 1500；学习率的初始值为 0.05；学习率的最终值为 0.01；邻域的初始值为 3；邻域的最终值为 0.5。

图 2 是 MViSOM 30×30 可视化实验结果，第一次标记的是会员编号为 464 的孤立点数据，第二次标记的是会员编号为 708 的孤立点数据，第三次标记的孤立点数据有四个，会员编号分别为 535，978，989，1071；第四次标记的孤立点数据有两个，会员标号为 302 与 938；第五次标记的孤立点数据有三个，会员编号为 76，148，1036；对于没有标记的孤立点数据 $x(t)$，假设其在输出空间的映射为 (i,j)，我们设置 $\sigma=2$，阈值 $f=4\cdot2^2\cdot0.25\cdot\frac{1082}{(30-1)\cdot(30-1)}\approx5.15$，如果以 (i,j) 为中心，以 2 为邻域宽度的区域数据点个数小于 5.15，则视数据 $x(t)$ 为稀疏点数据，并将其作为孤立点处理。表 1 中只列出图 2 中 1 到 4 次标记的孤立点数据，表 2 为整理出的会员年度乘机数据表中三个指标的平均值与标准差，表 3 为三个指标的相关系数列表。

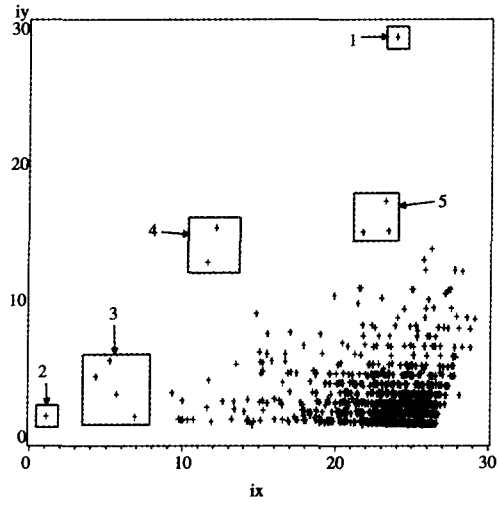

图 2　MViSOM 30×30 的可视化实验结果

表 1 孤立点挖掘结果列表

孤立点序号	会员编号	X1	X2	X3
1	464	51	48484	50774
2	708	11	56984	41891
3	535	14	56284	32381
4	978	15	52322	43462
5	989	11	44453	34373
6	1071	17	56191	40252
7	302	26	36946	55740
8	938	31	55304	45005

表 2 三个指标的平均值与标准差

	X1	X2	X3
平均值	13.392	13391.4	12999.33
标准差	3.285	7611.83	5831.29

表 3 各指标相关系数

	X1	X2	X3
X1	1.0000	0.3980	0.5390
X2	0.3980	1.0000	0.9238
X3	0.5390	0.9238	1.0000

从表 3 可以看出，各指标存在一定的相关性，有必要用独立成分分析的方法对原指标数据进行处理。从表 1 与表 2 可以看出，孤立点 1,7,8 的三个指标都很明显偏离中心值，孤立点 2,3,4,5,6,7 的指标 X2,X3 偏离中心值，与实际情况基本相符合。如果孤立点数据被确定，应当对其内容进行细查，根据其特征和挖掘目的而确定其是否为真正的孤立点数据。我们从该航空公司的会员信息表中查询，确认对表 1 中列出的八个会员为该航空公司的金卡会员，是重点客户，这也说明了 IMVOM 模型的合理性。

小结 一般高维数据的内部十分复杂，很难被探测，而人眼视觉对二维的几何空间分布却有着绝佳的分析能力，如果能够将高维数据资料转成人眼可以观察到的图形的话，对于高维数据的孤立点挖掘将是很有帮助的。本节在 ICA 与 MViSOM 的基础上提出了一个 IMVOM 孤立点挖掘模型，IMVOM 模型先通过 ICA 去除了数据之间的相关性，并通过 MViSOM 算法保持数据之间一定的拓扑结构不变性，并取得数据的可视化效果。最后，在孤立点挖掘的过程中，加入了“人为因素”。结合了“人类擅长于模式识别的能力”与“电脑擅长于大量地记忆、快速地计算的能力”的双方优点，IMVOM 方法避免了对高维数据内部结构的复杂探测，最终为克服高维数据集孤立点挖掘过程中的困难提供了基础。实验结果也验证了该模型的合理性，从而为数据挖掘提供了一种有效的方法。

参 考 文 献

1 Liu X. Strategies for outlier analysis. Birkbeck College University of London, 2000

2 Johanna H, Rocke D. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 2004, 44: 625～638

3 Bayarri M J, Morales J. Bayesian measures of surprise for outlier detection. Journal of Statistical Planning and Inference, 2003, 111: 3～22

4 Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms. Tsing hua University Press, 2003

5 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

6 Jutten C, Herault J. Independent component analysis versus PCA. In: Proceeding of European Signal Processing Conf. 1988. 287～314

7 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces. LNCS, 2001, 2234: 271～281

8 Kohonen T. Self-organizing maps. 3rd ed. Berlin Heidelberg New York: Springer, 2001

9 Yin H. ViSOM——a novel method for multivariate data projection and structure visualization. IEEE Transaction on Neural Networks, 2002, 1: 237～243

10 Yin H. Data visualization and manifold mapping using the ViSOM. Neural Networks, 2002, 15: 1005～1016

11 Sarvesvaran S, Yin H. Visualisation of Distributions and Clusters Using ViSOMs on Gene Expresssion Data. Lecture Notes in Computer Science, 2004, 3177: 78～84

12 Wu S, et al. PRSOM: A New Visualization Method by Hybridizing Multidimensional Scaling and Self-Organizing Map. IEEE Transactions on Neural Networks, 2005, 5: 1362～1380

13 彭红毅，蒋春福，朱思铭. 一种改进的高维数据可视化模型. 计算机科学(已录用)

14 彭红毅，蒋春福，朱思铭. 基于 ICA 与 SVM 的孤立点挖掘模型. 计算机科学，2006, 33(9)

(上接第 186 页)

6 Oliveria S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining. In: Workshop on Privacy, Security, and Data Mining at The 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December 2002

7 Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding Association Rules by Using Confidence and Support. In: Procedings of the 4th Information Hiding Workshop, 2001. 369～383

8 Pinkas B. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explorations, 2002, 4(2)

9 Chang Li Wu, Moskowitz I S. Parsimonious downgrading and decisions trees applied to the inference problem. In: Proceedings of the 1998 New Security Paradigms Workshop, 1998. 82～89

10 Evfimievski A, Srikant R, Agrawal R, et al. Privacy preserving mining of association rules. In: Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002. 217～228

11 Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining. In: Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, China, August 2002

12 Agrawal R, Srikant R. Privacy- preserving data mining. In: Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, 2000

13 秦静，张振峰，冯登国，等. 无信息泄露的比较协议. 软件学报，2004, 15(3)

14 Kantarcioglu M, Clifton C. Privacy- preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2002

15 Saygin Y, Verykios S, Elmagarmid A K. Privacy Preservin, Association Rule Mining. In: Procedings of the 12th Int'l Workshop on Research Issues in Data Engineering: Engineering e-Commerce/e-Bussiness Systems (RIDE'02)

16 Du Wenliang, Zhan Zhijun. Building decision tree classifier on private data. In: Procedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, 2002

17 Lindell Y, Pinkas B. Privacy preserving data mining. In: Advances in Cryptology-Crypto, 2000. 36～54

18 Yang Zhangqiang, Zhong Zhong, Wright R N. Privacy-preserving Classification of Customer Data Without Loss of Accuracy. In: Proceedings of the 5th SIAM International Conference on Data Mining, 2005. 21～23

19 Lin Xiaodong, Clifton C, Zhu M. Privacy-preserving clustering with distributed EM mixture modeling. Knowledge and Information Systems, 2004