Information Technology Journal 10 (4): 710-716, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.710.716

Survey of Data-mining Techniques used in Fraud Detection and Prevention

[1]Sheela Thiruvadi and [2]Sandip C. Patel
[1]Department of Accounting and Finance,
[2]Department of Information Science and Systems, Morgan State University, MD 21251, Baltimore, USA

Abstract: Data mining is a powerful tool widely used by organizations to enhance their businesses and gain a competitive advantage over their competitors. The data mining process helps in extracting and analyzing various data patterns, information or trends from large databases. Various data mining techniques are available to conduct the data mining process. Data mining techniques are used in a variety of applications, one of which is the detection and prevention of different types of frauds. Although there is existing research on data mining and various data mining techniques that can be used to detect and identify different types of frauds, there is little research that synthesizes various facets of fraud that uses the data mining techniques. In this survey study, we classify frauds into four categories with regards to the use of data mining as a tool in fraud detection and prevention. The four categories of fraud are management fraud, customer fraud, network fraud and computer-based fraud. We present the latest developments on the use of data mining as a tool for each of these categories.

Key words: Data mining, information security, intrusion detection, fraud prevention, network security

INTRODUCTION

Using the automated or semi-automated analysis, the data mining process extracts novel trends and patterns from large databases that can be used to make crucial business decisions. Data mining is a new concept that came into vogue in the 1990s and it is widely used in organizations to enhance their performance and gain a competitive edge over their competitors (Hormozi and Giles, 2004). The data mining process can be implemented using techniques such as decision trees, neural networks and Bayesian belief networks. Among its many different uses, data mining has been used as an effective tool in identifying frauds. Organizations lose substantial amount of their revenues due to fraudulent business practices, which can be defined as a deception devised for individual or collective gain using multifarious means such as unethical or illegal representation.

MANAGEMENT FRAUD

A fraud can be classified based on the party committing the fraud such as an employee, a customer, a vendor, an investor/consumer, or the company management (Albrecht *et al*., 2009). For example, in the management fraud the company's executive management is a culprit and they can involve in actions such as falsification or manipulations of expenses, invoices, sales figures, revenues, tax-liable items and so forth. The management could produce fraudulent data to the company stakeholders. More specifically, management fraud, often called financial statement fraud, is a deliberate and wrongful act carried out by public companies using materially misleading financial statements that may cause damage to investors, creditors and the economic market (Gill and Gupta, 2009). The research on management fraud could help auditors detect any wrongdoings in financial data by the executive management. Prior research has examined the association between a fraud and the audit-committee characteristics. For example, the in-depth analysis of such association by Huang and Thiruvadi (2010) showed that the presence of a financial expert on the audit committee helped in decreasing the likelihood of fraudulent financial reporting. Kirkos *et al*. (2007) used a sample of 76 Greek manufacturing companies in order to inquire and draw an analogy between the performance of the various factors that are associated with the financial-statement frauds. They used three data mining techniques, which were decision trees, neural networks and Bayesian belief networks. The authors used financial ratios of the balance sheets and income statements as the input data. The results showed that the data in the published financial statement contained falsified indicators similar to what prior research had indicated.

Corresponding Author: Sandip C. Patel, Department of Information Science and Systems, Morgan State University, MD 21251, Baltimore, USA Tel: (443) 885-3836 Fax: (443) 885-8255

Out of the three models used, the Bayesian belief networks performed the best with a 90.3% correct classification of the cross validation procedure, followed by the neural networks with an 80% success rate and finally the decision trees model with a 73.6% success rate. Although both prevention and detection of fraud should be considered by an organization for an effective anti-fraud policy, preventing a fraud is much more cost-effective than detecting it and it should take precedence (Bologna and Lindquist, 1995) in terms of resource allocation. Gill and Gupta (2009) used generic data mining framework for fraud prevention along with fraud risk-reduction for the financial-statement fraud. Data mining tasks were divided into two groups by the authors: predictive tasks and descriptive tasks. Predictive data mining, along with machine learning helped in better fraud prevention. The performance evaluation of various data mining techniques, as described above, was conducted by the authors using metrics such as error rate, information gain and Gini Index for decision trees. In another study (Huang *et al*., 2008), the aim of the study was to help auditors in identifying any possible fraud records and evaluating datasets by developing a fraud detection mechanism based on Zipf's Law through a simulation test and a case study. Four key performance indicators, Audit Hit Rate, Bayes Audit Hit Rate, Confusion Matrix and the misclassification cost matrix were used. Results showed that Zipf's mechanism could enable auditors reduce suspicion by identifying the differences between normal and abnormal data occurrences. The results showed that fraud record that had frequent sequential patters could be identified by Zipf Analysis and Zipf Analysis was more effective than a 100% sampling.

Since frauds and intrusions are frequently revealed by examining patterns of user activities, many fraud identification methods recognized misuse patterns that can classify a user's activity as normal or abnormal, which can suggest a possible fraud. Intelligent techniques such as and Support Vector Machines (SVM) and data mining are used to design fraud detection systems. New statistical models and methods help in enhancing fraud detection. For example, Kotsiantis *et al.* (2006) examined the efficiency of the machine learning techniques in identifying firms that publish Fraudulent Financial Statement (FFS) by implementing a hybrid decision support system through combining algorithms that uses a stacking variant methodology. The authors used a sample containing data from 164 non-financial Greek manufacturing firms listed on the Athens Stock Exchange (ASE), 41 of which had issued FFS. The authors collected the variables from financial statements of those firm. Results from the experiment corresponded to prior research, which indicated that the falsification indicators and a small list of ratios largely determined the classification results in published financial statements. Liou (2006) explored the similarities and differences between two models namely, fraudulent financial reporting detection and the Business Failure Prediction (BFP), that helped in identifying firms that procured losses to their investors, stakeholders and creditors. The purpose of this exploration of the two models was to find the effectiveness of the approach and the explanatory variable. The authors used three different data mining algorithms which were, logistic regression, neural networks and classification trees in order to construct the detection/prediction models using the data from Taiwan Economic Journal data bank and the Taiwan Stock Exchange Corporation website. The financial variables spanning the period of year 2003 to 2004 was used for calculations. Results showed that although many of the variables were significant in detecting fraudulent financial reporting and predicting business failure, logistic regression was considered the best of the three data mining algorithms specified above. Many computerized systems use electronic models that get triggered when certain business conditions arise to buy or sell stocks. Unauthorized triggering of such computerized financial models can be used for fraudulent or malicious purposes. In an interesting new research, Bapna and Patel (2010) proposed two frameworks to analyze the security of systems that used the computerized data-triggering models. The authors proposed a deterministic specification distributed intrusion detection system to secure the data-triggering model systems from internal as well as external threats. It is important to assess vulnerability of an information system in quantitative figures to provide idea as to the degree of security. Patel *et al.* (2008) proposed a novel approach for assessing an organization's vulnerability to information-security breaches using vulnerability trees. To help quantitatively measure the level of cyber-security for a computer-based information system, the authors presented two new indices, referred as the threat-impact index and the cyber-vulnerability index.

CUSTOMER FRAUD

In customer fraud, a customer either does not pay for the goods s/he purchased, or deceives organizations into giving them that they should not have (Albrecht *et al*., 2009). To detect a fraud, vast amounts of data from financial transactions, marketing surveys, medical records and health informatics are used. Data mining uses classification algorithms and learns from the related or linked data. If the number of examples in one class is

significantly different than the other, the fraud-detection method tends to produce poor predictive performance over the underrepresented class. Guo and Viktor (2008) discussed the use of a new strategy to address the imbalance in multi-relational data wherein the one class in the target relation is higher than the others. Such imbalances help in different discoveries, such as diagnosing a disease or detecting a fraud case, for example, a credit-card fraud. Experiments were performed using six benchmark data-sets. The results indicated that the suggested method was better than other prevailing data mining algorithms in comparison, especially when there was a high class-imbalance with regards to ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve). Similarly, Nonyelum and Chibueze (2009) used the neural network technology and the rule-based component to develop credit-card fraud-detection (CCF) system using four clusters (low, high, risk and high risk), in lieu of the two-stage model that is frequently used in fraud detection algorithms. A model identifying the behavior of a cardholder and evaluating the transaction characteristic to detect fraudulent transactions was developed using the self-organizing-map algorithm. Several models were generated by applying the artificial neural network trained with the unsupervised learning methods. This generation was done to secure a correct result and minimize the wrongful classification in which a genuine transaction is considered fraudulent. Interesting fraud detection models and frameworks have also been suggested recently. A fraud detection framework by Xu *et al.* (2007) used data mining algorithm on simulated and real data to create user profiles for identifying customer behaviors in detecting fraudulent transactions in an online system through a set of association rules. Anomalies were identified by comparing the incoming transactions of the user against that user's profile based on his/her recent transactions. Results from the experiments showed that the differences between the anomaly behavior and the profiled user behavior can be correctly interpreted by the proposed algorithm.

NETWORK FRAUD

The telecommunication and other networks play a major role in allowing/preventing and detecting frauds. In the telecommunication fraud, either the network services are the target of fraud, or they are used as a means to commit other forms of fraud. Identity thefts, telemarketing frauds, investment scams, skimming or money-offer schemes are some examples of telecommunication frauds. Network access by unauthorized intruders can compromise the system, making the system vulnerable to frauds and other security breaches (Graham and Patel, 2006). Classification of network traffic helps identify abnormal behaviors by detecting any deviations from the normal activities. For example, Kou *et al.* (2009) discussed data mining-based network intrusion detection system and tacked the problem of solving the multi-class classification, in which network traffic is classified into more than two groups. These authors introduced a model referred to as the multi-criteria mathematical programming model for multi-class classification for the detection of network intrusion and defined the e-support vector for reducing complexity in a large-scale application. Two network intrusion datasets were suggested as performance test for better accuracy and lower false alarms for the intrusion classification. The datasets used were KDD99, which was a four-class dataset and NeWT, which was a three-class dataset from the University of Nebraska, trained with classification algorithms.

A fraud can also be committed using networking or telecommunication security-breaches. Becker *et al.* (2010) argued that some of the different frauds in telecommunications are (1) subscription fraud, (2) intrusion fraud, (3) fraud based on loopholes in technology, (4) social engineering, (5) fraud based on new technology, (6) fraud based on new regulation and (7) masquerading as another user. Becker *et al.* (2010) discussed different strategies and techniques used in the detection of the telecommunication-fraud history at AT and T Inc. A fraud-management system was developed to manage different types of frauds. Different key elements were discussed by the authors to detect frauds, which included call details, database required for storing data, fraud-detection algorithms, fraud types and corrections and visualization tools that can help in diagnosis. The authors further discussed how the detection of fraud has moved from the early threshold-based alerting to graph based signatures and stressed that humans should be involved in fraud detection rather than depending only on such tools.

Time is of essence when it comes to finding a fraud. The real-time detection could save a significant amount of money by preventing the fraud from occurring in the first place. Using the probabilistic models could improve the speed of fraud detection. Xing and Girolami (2007), for example, addressed the issue of the telecommunications fraud by building user profile signatures using Latent Dirichlet Allocation (LDA) and presupposed that any deviations from the normal behavior of individual users was highly correlated with fraudulent activity. The LDA could efficiently infer user calling behavior by joining the different classes of distribution. In order to score calls, comparisons were made between a normal user making a call versus a fraudster making the call. Results showed that the use of probability distribution, combined with

that of the fraudster activity would improve the chances of detecting the fraudulent calls. Further analysis showed that this experiment was effective, precise and could help in real-time fraud detection. Recent social and business trends such as globalization and use of the Internet for social networking have some researchers investigate into newer, more applicable fraud detection methods. For example, a recent work by Liao *et al*. (2009) discussed the need for an effective and automated system for network forensics as compared to the earlier tools and methods that were not very effective when there was increased network traffic. The authors suggested an approach that could examine computer crimes in a networked atmosphere and automatically create digital evidence using fuzzy logic and expert system. The experimental results indicated that 91.5% of the attack types could be classified by the system thereby providing understandable information for forensic experts.

We observed that different data mining techniques have different success rates. A study by Sanver and Karahoca (2009) compared different data mining techniques, benchmarked each technique and identified Adaptive Neuro Fuzzy Inference (ANFIS) for telecom-fraud detection in Turkey. The results showed that ANFIS provided 97% of sensitivity, 99% of specificity, where 98.33% of the instances were correctly classified. Different data mining methods such as JRip, PART, Ridor, OneR, Nnge, Decision Table, Conjunction Rules, AD Trees, IB1 and Bayesian networks were used and compared. Results showed that ANFIS classified data yielded more precise results by joining the accuracy of fuzzy-logic-based classification system and the adaptable feature of neural networks, particularly when there was a concern for false positive rate. Additionally, the results showed that the ANFIS model could successfully work as a model to detect a fraud. Izhan *et al*. (2009) discussed the use of the Gaussian Mixed Models (GMM) as a probabilistic data mining model. This model was used for the detection of fraud, which identified the voice of a person through the speech recognition based on the number and duration of calls taking place during all hours, both nationally and internationally. The maximum likelihood estimation for GMM was found with the help of the Expectation Maximization (EM) algorithm. The authors used a simulated program using the Box and Muller transformation for generating the simulation data. Although different combinations of parameters and components were used by inspecting histograms and plots to start the EM algorithm, plotting was used as the graphical technique since the plotting helped in determining the EM algorithm faster and easier. In addition to the benefits listed above for using neural network in conjunction with other methods, a particular type of neural network has also been shown to yield success in intrusion detection and fraud protection. For example, using neural network architecture referred as the Back-Propagation Neural (BPN) Network, Om and Sarkar (2010) discussed the problem of unauthorized access to web documents that do not have any value to a company but prove to be harmful for the organization once retrieved by the unauthorized Internet intruders. By using the web proxy requests, such documents in an organization were evaluated and monitored through the use of Microsoft Internet Security and Acceleration server 2000. Various steps such as (1) selected IP address converted to an integer number without delimiter, (2) input pattern normalized into real number, (3) BPN trained for input pattern taking different number of different layers and (4) BPN tested for test patterns after training, were performed for a valid input for BPN. The results showed that this method was useful for blocking unauthorized access to the worthless web documents and saving network bandwidth by identifying such documents.

As listed earlier, the data mining process largely depends on the classification of data and the abstracting rules and knowledge from the data patterns. A better rule-extraction method results in improving fraud detection. SVM work well with the decision making and has been used successfully in cell-phone fraud detection. Mukkamala *et al*. (2002) found SVMs to be superior to artificial neural networks in many important respects of intrusion detection. Pang and Kasabov (2009) proposed a new method for rule extraction, referred to as the constructive method. In this method, the authors encoded the knowledge of data into SVM classification tree (SVMT) and extracted the linguistic association rule through decoding SVMT using unclear numerical parameters and dataset class-imbalance. This new rule made the decision-tree rule efficient, while keeping the SVM accuracy. The authors tested the proposed method on three applications, which include the cell-phone fraud detection using a database obtained from a mobile telecom company, cancer diagnoses and the Gaussian synthetic data. The results indicated that the SVMT-rule was accurate for data classification and was superior as compared to a purely support-vector based rule extraction.

In a research related to the Next Generation Networks (NGNs), Bihina Bella *et al*. (2009) suggested the convergence to the NGNs, based on the Internet protocol with the original architecture for the Fraud Management System (FMS). This convergence could help in effective detection of frauds in NGNs due to the flexible and application-independent design, network coverage and scalability of the original architecture in lieu of the FMS.

The research-design contained seven components that passed through a four-stage detection operation, out of which the Self-Organizing Maps (SOMs) helped in the NGN fraud detection. The efficiency of the SOM in fraud detection was tested using a prototype tool that helped in further identifying usage patterns and outliers. In addition, the prototype helped in defining new fraud scenarios, which proved that the proposed FMS architecture worked effectively.

Outlier detection is a fundamental issue in data mining, specifically in fraud detection, network intrusion detection and network monitoring. Yamanishi *et al.* (2004) suggested the SmartSifter detection program, which is online unsupervised outlier-detection. The experiment using simulation data, network intrusion detection data and other data provided by the Australian Health Insurance Commission were used to demonstrate the detection program's function, its accuracy and the computational time required. This framework could be applied to other data mining tasks in addition to fraud detection. New research on mathematical approaches helps data mining techniques get more efficient. For example, outliers play an interesting role in complex data mining processes and help predict frauds such as the credit card fraud. Outliers could be ignored as noise (border object) or as outstanding/distinctive outliers. Yu *et al.* (2006) presented a new technique referred as the mutual-reinforcement-based (k-LOF) local outlier detection approach to help in identifying local outliers in the center compared to the current method of density-based (LOFk) that identifies outliers as noise. In addition to identifying the outliers in the center, this approach makes an effort to treat them as numerical data. This technique could reduce the burden on users by ascertaining the relationship between data items and find reasons for the occurrence of outliers.

COMPUTER-BASED FRAUD

Computer-security loopholes can provide a means to commit fraud to a fraudster. People committing frauds take advantage of computer security flaws. We examined research on security attacks and the use of data mining techniques to detect them. Kolter and Maloof (2006) used machine learning and data mining to discover and classify malicious executables. The research selected executable which would appear undetected on a user's hard drive, without preprocessing or removing any obfuscation. Benign and malicious executables were encoded using n-grams. A variety of inductive methods such as naive Bayes, decision trees, support vector machines and boosting were determined after the selection of related n-grams. The results showed that the boosted decision trees had an area under the ROC curve of 0.996, surpassing other models. The three major contributions of the authors' work were: (1) detected and classified malicious executables using the established methods of text classification, (2) presented empirical analysis for detecting and classifying malicious executables in the wild using inductive methods and (3) achieved high detection-rates even on new malicious executables. Likewise, Mukkamala *et al.* (2005) showed that the ensemble of artificial neural network, SVM and Multivariate Adaptive Regression Splines, was superior to individual approaches for intrusion detection in terms of classification accuracy. The authors used data from Massachusetts Institute of Technology's Lincoln Lab and classified five different classes of patterns in the data from Defense Advanced Research Projects Agency. The results also showed that 100% classification accuracies can be achieved if appropriate intelligent paradigms are chosen. New research, such as that by Hua *et al.* (2009), proposed practical approaches for selecting and implementing organizational Information System Security (ISS). Hua *et al.* (2009) presented three models for securing business information systems, namely, ISS offense model, ISS defense model and the safeguard model. The offence and defense model illustrated the objects that help breach system security and prevent the breaches, respectively. The safeguard model examined the relationship among three security aspects, namely, business-security requirement, security techniques and the controls that are available for security enhancements. Organizations need to estimate the loss from a potential security breach to their information systems. In a recent research, Patel and Zaveri (2010) proposed a risk-assessment model to assess the financial damages resulting from these cyber attacks. Attempts have also been made to prevent attacks or safeguard information from a specific type of information system such as a critical infrastructure information systems (Patel *et al.*, 2009), or the information system that use radio frequency identification (Patel and Emdad, 2008), which are new and emerging type of information systems with much privacy concerns.

CONCLUSION

In this study, we have discussed effective uses of different data mining techniques for detecting and preventing fraudulent activities and categorized fraud. Specifically, we discussed data mining techniques as it relates to four classes of frauds, which are management fraud, customer fraud, network fraud and computer-based

fraud. We presented the latest developments in these areas. This study would provide guidelines to business practitioners who want to employ data mining techniques to detect fraud in their companies, auditors who would like to identify frauds and the computer administrators, who need to safeguard their computers and networks from fraudulent activities. This study would also provide direction to the researchers who would like to explore more effective techniques of detecting and preventing frauds, or extending data mining research for its more effective use.

REFERENCES

Albrecht, W.S., C.C. Albrecht, C.O. Albrecht and M.F. Zimbelman, 2009. Fraud Examination. South-Wester Cengage Learning, Mason, Ohio.

Bapna, S. and S. Patel, 2010. Securing computerized models and data against integrity attacks. Int. J. Electronic Finance, 4: 343-354.

Becker, R.A., C. Volinsky and A.R. Wilks, 2010. Fraud detection in telecommunications: History and lessons learned. Technometrics, 52: 20-33.

Bella, M.A.B., J.H.P. Eloff and M.S. Olivier, 2009. A fraud management system architecture for next-generation networks. Forensic Sci. Int., 185: 51-58.

Bologna, G.J. and R.J. Linquist, 1995. Fraud Auditing and Forensic Accounting. John Wiley and Sons Inc., New York.

Gill, N.S. and R. Gupta, 2009. Prevention and detection of financial statement fraud: A data mining approach. IUP J. Syst. Manage., 7: 55-68.

Graham, J. and S. Patel, 2006. Internet-based secure monitoring and control for utility companies and process plants. DIAS Technol. Rev. Int. J. Bus. IT, 3: 28-33.

Guo, H. and H.L. Viktor, 2008. Learning from skewed class multi-relation databases. Fundamenta Informaticae, 89: 69-94.

Hormozi, A.M. and S. Giles, 2004. Data mining: A competitive weapon for banking and retail industries. Inform. Syst. Manage., 21: 62-71.

Hua, J., S. Patel and J. Zaveri, 2009. Securing business information systems from cyber-attacks. J. Digital Bus., 3: 35-53.

Huang, S.M., D.C. Yen, L.W. Yang and J.S. Hua, 2008. An investigation of Zipf's Law for fraud detection (DSS#06-10-1826R(2)). Decision Support Syst., 46: 70-83.

Huang, H.W. and S. Thiruvadi, 2010. Audit committee characteristics and corporate fraud. Int. J. Public Information Syst., 6: 71-82.

Izhan, M., M. Yusoff, M.R. Abu-Bakar and A.H.S. Mohd Nor, 2009. The performance of Expectation Maximization (EM) algorithm in Gaussian Mixed Models (GMM). Perlanika J. Sci. Technol., 17: 231-243.

Kirkos, E., C. Spathis and Y. Manolopoulos, 2007. Data mining techniques for the detection of fraudulent financial statements. Expert Syst. Appl., 32: 995-1003.

Kolter, J.Z. and M.A. Maloof, 2006. Learning to detect and classify malicious executables in the wild. J. Machine Learning Res., 7: 2712-2744.

Kotsiantis, S., E. Koumanakos, D. Tzelepis and V. Tampakas, 2006. Forecasting fraudulent financial statements using data mining. Int. J. Computational Intell., 3: 104-110.

Kou, G., Y. Peng, Z. Chen and Y. Shi, 2009. Multiple criteria mathematical programming for multi-class classification and application in network intrusion detection. Inform. Sci., 179: 371-381.

Liao, N., S. Tian and T. Wang, 2009. Network forensics based on fuzzy logic and expert system. Comput. Commun., 32: 1881-1892.

Liou, F.M., 2006. Fraudulent financial reporting detection and business failure prediction models: A comparison. Manage. Auditing J., 23: 650-662.

Mukkamala, S., A.H. Sung and A. Abraham, 2005. Intrusion detection using an ensemble of intelligent paradigms. J. Network Comput. Appl., 28: 167-182.

Mukkamala, S., G. Janoski and A. Sung, 2002. Intrusion detection using neural networks and support vector machines. Proceedings of IEEE International Joint Conference on Neural Network, May 12-17, Honolulu, HI, USA., pp: 1702-1707.

Nonyelum, O.F. and I.H. Chibueze, 2009. Credit card fraud detection using artificial neural networks with a rule-based component. ICFAI Univ. J. Sci. Technol., 5: 40-47.

Om, H. and T.K. Sarkar, 2010. Designing intrusion detection system for web document using neural network. Commun. Network, 2: 54-61.

Pang, S. and N. Kasabov, 2009. Encoding and decoding the knowledge of association rules over SVM classification trees. Knowledge Inform. Syst., 19: 79-105.

Patel, S. and A. Emdad, 2008. Security considerations for RFID in supply chain management. Academy Taiwan Bus. Manage. Rev., 4: 12-19.

Patel, S., G. Bhatt and J. Graham, 2009. Improving the cyber security of SCADA communication networks. Commun. ACM, 52: 139-142.

Patel, S. and J. Zaveri, 2010. A risk assessment model for cyber attacks on information systems. J. Comput., 5: 352-359.

Patel, S., J. Graham and P. Ralston, 2008. Quantitatively assessing the vulnerability of critical information systems: A new method for evaluating security enhancements. Int. J. Inform. Manage., 28: 483-491.

Sanver, M. and A. Karahoca, 2009. Using Fraud detection using an adaptive neuro-fuzzy inference system in mobile telecommunication networks. J. Multiple-Valued logic Soft Comput., 155: 155-179.

Xing, D. and M. Girolami, 2007. Employing latent dirichlet allocation for fraud detection in telecommunications. Pattern Recognition Lett., 28: 1727-1734.

Xu, J., A.H. Sung and Q. Liu, 2007. Behavior mining for fraud detection. J. Res. Practice Inform. Technol., 39: 3-18.

Yamanishi, K., J. Takeuchi, G. Williams and P. Milne, 2004. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining Knowledge Discovery, 8: 275-300.

Yu, J.X., W. Qian, H. Lu and A. Zhou, 2006. Finding centric local outliers in categorical/numerical spaces. Knowledge Inform. Syst., 9: 309-338.