The International Journal of Digital Accounting Research
Vol. 11, 2011, pp. 69 - 84
ISSN: 1577-8517

Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach[1]

Sutapat Thiprungsri. Rutgers University. USA. sutapat@pegasus.rutgers.edu

Miklos A. Vasarhelyi. Rutgers University. USA. miklosv@andromeda.rutgers.edu

Abstract This study examines the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. Cluster analysis groups data so that points within a single group or cluster are similar to one another and distinct from points in other clusters. Clustering has been shown to be a good candidate for anomaly detection. The purpose of this study is to examine the use of clustering technology to automate fraud filtering during an audit. We use cluster analysis to help auditors focus their efforts when evaluating group life insurance claims. Claims with similar characteristics have been grouped together and small-population clusters have been flagged for further investigation. Some dominant characteristics of those clusters which have been flagged are large beneficiary payment, large interest payment amounts, and long lag between submission and payment.

Keywords: Continuous auditing, cluster analysis, anomaly detection, insurance industry.

1. INTRODUCTION

This study examines the application of cluster analysis in the accounting domain, in particular its application to discrepancy detection in the field of audit. Clustering is an unsupervised learning algorithm, which means that data are analyzed without the presence of predetermined labels (e.g. "fraudulent/non-

[1] We would like to thank Mr. JP Krahel, the participants of the Rutgers CarLab workshops and many other colleagues for their many suggestions.

DOI: 10.4192/1577-8517-v11_4

Submitted June 2011
Accepted July 2011

fraudulent") (Kachigan, 1991). Clustering is a technique for grouping data points so that points within a single group (or "cluster") are similar, while points in different clusters are dissimilar. As an unsupervised learning algorithm, clustering is a good candidate for fraud and anomaly detection techniques because it is often difficult to identify abnormal / suspicious transactions. Clustering can be used to group transactions so that different levels of attention and effort can be applied to each cluster. The purpose of this study is to apply clustering techniques to the audit field. Automated fraud filtering can be of great value as a preventive tool (Vasarhelyi *et al.*, 2011). We apply cluster analysis to a unique dataset provided by a major US insurance company and examine the resulting outliers. Cluster-based outliers help auditors focus their efforts when evaluating group life insurance claims. Some dominant characteristics of outlier clusters are large beneficiary payment, large interest payment amounts, and long lag between submission and payment.

In the next section, we discuss the problem, followed by a review of the relevant multidimensional clustering literature is reviewed. We then explain the insurance claims setting, present results, and conclude.

2. LITERATURE REVIEW

2.1 Anomaly and Anomaly Detection

Outliers are observations that deviate so much from other observations that they may have been generated by a different mechanism (Hawkins, 1980). Anomalies occur for many reasons. For example, data may come from different classes, there may be natural variation in data or measurement, or collection error may have occurred (Tang *et al.*, 2006).

Chandola *et al.* (2009) suggest that anomalies can be classified into three categories: 1) point anomalies, 2) contextual anomalies, and 3) collective anomalies. A point anomaly is an individual data instance which is identified as anomalous with respect to the rest of the data. A contextual anomaly occurs when a data instance is anomalous in a specific context. For example, a temperature of 35°F is considered normal in winter, but anomalous in summer (Chandola *et al.*, 2009). Collective anomaly occurs when a collection of related data instances is

anomalous. Duan *et al.* (2009) suggests that abnormal events may exhibit both temporal and spatial locality, forming small outlier clusters. This phenomenon is called a "cluster-based outlier".

Anomaly detection is the task of identifying observations with characteristics that significantly differ from the rest of the data (Tang *et al.*, 2006). Applications of anomaly detection include fraud, credit card fraud, network intrusion, to name a few. Regardless of domain, anomaly detection generally involves three basic steps: 1) identifying normality by calculating some "signature" of the data, 2) determining some metric to calculate an observation's degree of deviation from the signature, and 3) setting thresholds which, if exceeded, mark an observation as anomalous (Davidson, 2002). A variety of methods for each step has been used in many fields.

Chandola *et al.* (2009) suggest that, with respect to label availability, anomaly detection can operate in one of three modes: 1) supervised, 2) semi-supervised, and 3) unsupervised. Supervised anomaly detection assumes the availability of a training data set which has instances labeled as normal or anomalous. Semi-supervised anomaly detection assumes that the training data set includes only normal instances. A model corresponding to normal behavior will be built and used to identify anomalous instances in the test data. Unsupervised anomaly detection does not require any training dataset, instead simply assuming far fewer anomalies than normal instances.

2.2 Cluster Analysis for Anomaly Detection

Chandola *et al.* (2009) propose that clustering based techniques for anomaly detection can be grouped into three categories:

1. The first group assumes that normal instances belong to a cluster while anomalies do not belong to any cluster. Examples include DBSCAN-Density-Based Spatial Clustering of Applications with Noise (Ester *et al.*, 1996), ROCK-Robust Clustering using links (Guha *et al.*, 2000), SNN cluster-Shared Nearest Neighbor Clustering (Ertoz *et al.*, 2003), FindOut algorithm (Yu *et al.*, 2002) and WaveCluster algorithm (Sheik-holeslami *et al.*, 1998).

These techniques apply a clustering algorithm to the data set and identify instances that do not belong to a cluster as anomalous.

2. The second group assumes that normal data instances lie closer to the nearest cluster *centroid* (or center) while anomalies are far away from the nearest cluster *centroid* (Chandola *et al.*, 2009). Self-Organizing Maps (SOM) introduced by Kohonen (1997) are used for anomaly detection in many different applications, including fraud detection (Brockett *et al.*, 1998) and network intrusion detection (Labib *et al.*, 2002, Smith *et al.*, 2002, Ramadas *et al.*, 2003). The techniques in this group involve two steps: grouping data into clusters, and calculating distances from cluster *centroid*s to identify anomaly scores. Chen *et al.* (2007) use local outlier factor (LOF) values to measure the outlying behavior among peer groups to gauge the financial performance of companies.

3. The third group assumes that normal data instances belong to large, dense clusters, while anomalies belong to small or sparse clusters (Chandola *et al.*, 2009). He *et al.* (2003) propose a technique called FindCBLOF to determine the size of the clusters and the distance between an instance and the nearest cluster *centroid*. Combining these two values return the Cluster-Based Local Outlier Factor (CBLOF). Applying the technique to detect anomalies in astronomical data, Chaudhary *et al.* (2002) propose an anomaly detection model using k-d trees (k dimensional tree – a k-dimensional space partitioning data structure for optimizing points) providing partitions of data in linear time. Sun *et al.* (2004) propose a technique called CD-trees. Both techniques define sparse clusters as anomalies.

3. THE SETTING: INSURANCE GROUP CLAIMS

This study focuses on data from the group life claims business of a major US insurance company. Group life claims is a type of group insurance marketed to corporate clients, typically covering most of their employees, often with different levels of available or required coverage.

Group life insurance is sold to companies in volume. For example, Company A buys group life insurance for 100 employees; while individual life insurance is

sold individually to John Doe. From the perspective of the insurance provider, the purchasing company is the customer in the former case; while the insured is the customer in the latter. The insurance company manages policies and claims from these two types differently. Importantly for our case, group insurance providers will not keep the information on insured individuals as individual providers do. Information regarding a particular insured employee is collected and entered when a claim is received.

The nature of group life insurance carries many risks for policy administration and audit. Very little is known about the type of risks or possible fraudulent actions within the group life claim insurance. The fraudulent claims previously found by internal auditors were happenstance and/or through hotline. The types of related risks identified by internal auditors are, for example, duplicate payments, fictitious names, improper/incorrect information entered into the systems, and suspicious payment amounts. Several tests are performed to check for the reasonableness of the amount of payments. The issues that ensue were drawn from a series of studies performed with the insurance files of several companies.

Current internal audit tests in the insurance context are, unfortunately, not very effective. For example, auditors check for fictitious insured names or fictitious beneficiary names. However, the definition of "fictitious name" is not clear; internal auditors must use judgment to identify suspected fictitious names. Manual data collection hinders data quality and test usefulness. For example, –calculating the age from birth date, death date and hiring date is not possible. Given a wrong dates of birth and death for an individual, resulting ages could be invalid. For example, there are several cases in which an insured's age, derived from a calculation ((Death date-Birth date)/365), is one digit (0-9 years old). The result would be normal if the data is for normal life insurance. However, an insured cannot be younger than the legal working age for group life insurance. Internal auditors are well aware of the problems and the shortcoming in their audits and are seeking innovative methods to control and reduce the risk of fraudulent claims. The purpose of this study is to apply the use of cluster technology to the internal audit process.

3.1 Data

The dataset contains the records of group life claim payments issued in the first quarter of 2009. The data include 208 attributes related to group life claims under one of five basic categories:

- Insured
- Coverage
- Group / company
- Beneficiary
- Payment

Clients submit a paper claim document to the insurance company. The claim document is then scanned and saved as a PDF file. The information will be manually entered into a verification system. Because the data is manually entered, several mistakes can be found in the data, some intentional and others due to carelessness. There are other cases in which the insured's death date is prior to birth or employment date. However, some mistakes are not so easily identified. For example, by using the data, a calculated age of the insured can be reasonable misleading, For example establishing a much earlier age than retirement age. Some may argue that the retirement date can be used to estimate the age and/or birth date. However, one individual can retire at age 60; while another retires at 50. It is impossible to identify the exact age when the insured decided to retire from the job. In the insurance industry, data is often old and entered over many years into many different legacy systems that were kept to support particular types of policies. With the passage of time, fields can be reused and the quality of edits and the underlying data deteriorates.

The sample contains 40,080 group life insurance claims paid out in the first quarter of 2009. After multiple examinations of the raw data and consultation with the internal auditor (a presumable domain expert), suggestions on attribute selection were given. Based on current internal audit control tests, payments are the major concern of this business unit. The value of the payment that the beneficiary initially receives is called the beneficiary payment. Depending on the state where the insured lives, the location of the company, and/or the where the beneficiary lives, if the payment is not made immediately after the claim is

received/approved, interest may accumulate. If so, the total payment that the beneficiary receives will consist of the beneficiary payment and interest payment. Several existing tests relate to reasonableness of interest and beneficiary payments. The internal auditors have expressed an interest in better and more useful techniques to test the reasonableness of the values.

Due to aforementioned data quality issues, and after consultation with expert internal auditors, a new dataset was created based on the original data. Two newly created attributes were selected for clustering:

- Percentage: Total interest payment / Total beneficiary payment
- AverageDTH_PMT: Average number of days between the death date and the payment date (a weighted average is used if a claim has multiple payment dates)

These attributes were normalized for comparison, reducing the impact of scale differences.

3.2 Clustering Procedure

Because all attributes are numeric, we used a simple K-mean clustering procedure[2]. K-mean clustering is a simple, well-known algorithm. It is less computer-intensive than many other algorithms, and therefore it is a preferable choice when the dataset is large (Tang *et al.*, 2006). The steps in K-mean clustering are as follows (Roiger *et al.*, 2003):

1. Choose a value for K, the total number of clusters to be determined.
2. Choose K instances (data points) within the dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign to remaining instances to their closet cluster center.
4. Use the instances in each cluster to calculate a new mean for each cluster.

[2] The software used for this analysis includes SAS and Weka. The data set was cleaned and transformed using SAS statistical package, exported into a comma separated value (CSV) file, and converted into ARFF (Attribute-Relation File Format) before being fed into Weka.

5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

Several numbers of clusters are tested. The number of cluster is chosen so that adding another cluster will not create a significantly improved model. The stopping point is determined by checking the percentage of variance explained as a function of number of clusters. When the first few clusters are added, they will add much explanation of variances. At some point, the marginal gain in variance explained will be reduced. The ideal number of clusters chosen should be at the point where the marginal gain begins to fall.

3.3 Anomaly Detection

Instead of selecting clustering techniques which require extensive programming, a simple clustering technique which can be easily performed on open source software, WEKA[3], is used. The purpose of this paper is to provide an example of how cluster analysis can be used for auditing. We chose a technique that is easy to use and interpret. This research assumes that both individual observations and small clusters can be outliers. Most points in the dataset should not be outliers. Outliers are identified in two ways. First, observations that have low probability of being a member of a cluster (i.e. are far away from other cluster members) are identified as outliers. The probability of 0.6 is used as a cut-off point. Second, clusters with small populations (less than 1% of the total) are considered outliers.

4. RESULTS

Because of the simplicity and suitability of the techniques to the data type, simple K-mean has been used as the clustering procedure. The 40,080 claims which are paid in the first quarter of 2009 are used in the analysis. The number of clusters selected is eight.

Enhanced results from Weka are shown in Table 1.

[3] WEKA(Waikato Environment for Knowledge Analysis) is an open source software develop by University of Weikato, New Zealand. It is freeware and offers functionality for many machine learning techniques; for example, various clustering techniques (such as DBSCAN, K-Mean, Cobweb, and etc), Decision Tree, Bayesian Network, Support Vector Machine, etc.).

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 8 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     TestSetPayment2
Instances:    40080
Attributes:   3
              N_AverageDTH_PMT
              N_percentage
Ignored:
              CLM_ID
Test mode:    evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 55
Within cluster sum of squared errors: 3.9256036521001687
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute               Full Data          0          1          2          3          4          5          6          7
                        (40080)       (2523)       (54)       (84)      (222)      (295)       (31)      (768)    (36103)
N_AverageDTH_PMT               0       0.6374     15.177     3.5419     6.9858     0.8778    10.9006     2.7806    -0.1937
N_percentage                   0       0.2666     1.8334     9.3405     0.5042     3.4637    26.6913     0.3185    -0.1057

Clustered Instances

0      2523 (  6%)
1        54 (  0%)
2        84 (  0%)
3       222 (  1%)
4       295 (  1%)
5        31 (  0%)
6       768 (  2%)
7     36103 ( 90%)
```

Clusters with small number of population

Table 1: Clustering result with 2 variables

Using two attributes, eight clusters are formed. About 90% of claims are grouped into cluster 7 and 6% are in cluster 0 (Table 1). Three clusters (1, 2, and 5) have membership of less than 1% (54, 84 and 31 members, respectively). Examining the characteristics of these less populated clusters, we discover some unusual characteristics. Claims in these clusters have high interest/beneficiary payment percentage and/or claims with long period of time from death dates to payment dates.

- Claims in cluster 5 have high interest/beneficiary payment percentage and a long period between the death dates and the payment date. Cluster 1 claims have long period from death to payment dates.
- Claims in cluster 2 have high interest/beneficiary payment.

The total number of claims identified as possible anomalies from cluster-based outliers is 169. In addition to identifying small clusters, the probability of individual observations' cluster membership is examined. The claims, which have lower than 0.6 probabilities of belonging to the cluster they are assigned to, are identified as possible anomalies. 568 claims fit this criterion.

The visualized results are shown in Figure 1 and Figure **2** (with clusters marked).

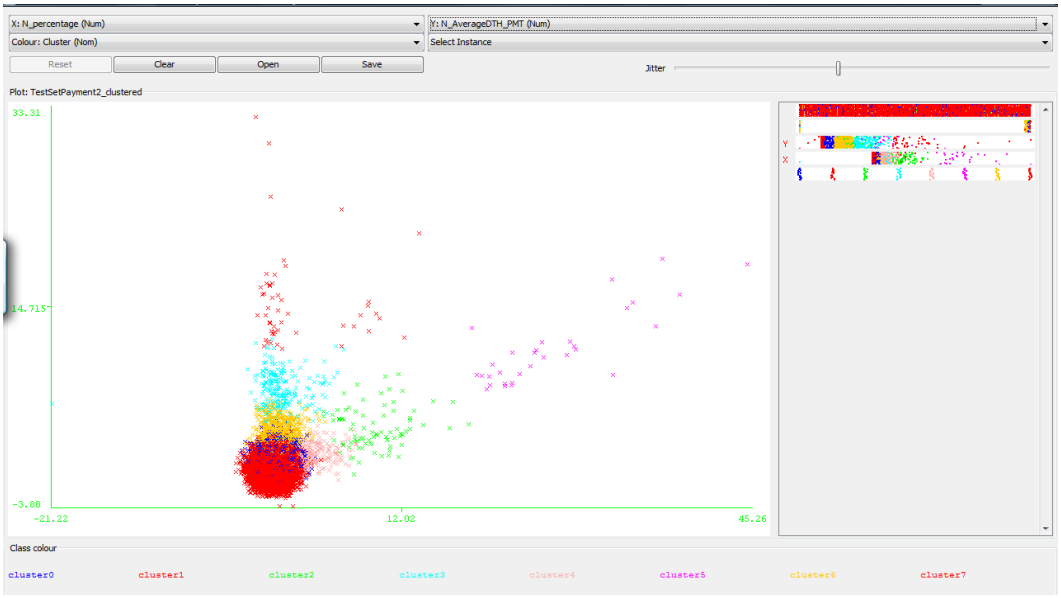

Figure 1: Visualization of the cluster assignment for 2 attributes clustering; N_Percentage and N_AverageDTH_PMT

Clusters 1, 2 and 5 are less (and more sparsely) populated. The characteristics of claims in these clusters are different from the majority of claims in other clusters. Having different characteristics does not necessarily signify abnormality or fraud. There are possible legitimate reasons; for example, high interest may be due to claim age. If the insured died long before a claim was submitted,

accumulated interest would be high.

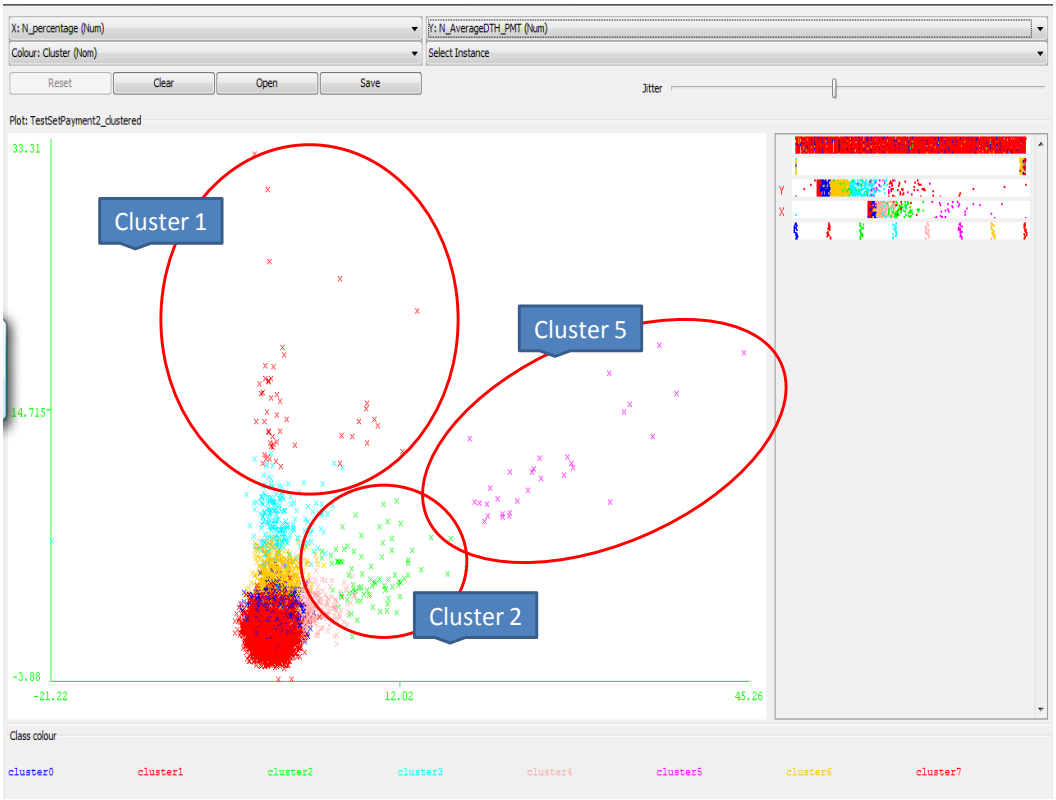

Figure 2: Visualization of the clustering result (two attributes) with cluster marked.

Giving that legitimate explanations are possible, it is wise to focus checks on those claims where non-fraudulent explanations are not available. For example, if a claim has a very high interest but the claim is young and the insured is only recently deceased, the internal auditor must investigate why interest is high. With a small number of claims identified, the internal auditor may also be free to perform additional calculations to check if the amount of interest is reasonable given the length of time the claim is present in the system and the length of time since the insured passed away.

To verify if cluster analysis can identify anomalies in accounting systems, suspicious cluster/individual claims should be selected for the further investigation by the internal auditor. Follow-up results would improve the model.

5. CLUSTER ANALYSIS AND ITS APPLICATION IN AUDITING

In addition to identifying groups within a dataset, clustering can be used to identify suspicious transactions or observations. Flagged transactions differ significantly from their peers. Transactions can be flagged due to extreme values, either low or high. These transactions can result from unintentional error or possible fraud. Further investigations will ideally distinguish between the two. Due to cost concerns (Cleary *et al.*, 2005), it is impossible for internal auditors to investigate all the flagged transaction. Therefore, decisions on materiality will have to be made concerning which flagged transactions should be pursued. Immaterial suspicious transactions may be left aside.

With the increasing complexity of transaction systems, fraudsters have new opportunities to commit fraud and outsmart the system. Auditors must seek new and innovative audit methods. Cluster analysis may flag transactions not identified via other methodologies. While universal detection is never guaranteed, flagged transactions demonstrate suspicious characteristics worth investigating.

Feedback from internal auditors can be very useful in improving the model. Whether flagged transactions end up being errors, fraud, or normal transactions, the validation of the results will provide inside knowledge which may be useful for the improvement of the model. Clustering results also provide more insight into the nature of the transactions by placing alike transactions into group and pointing out heterogeneities' in alike transactions.

6. CONCLUSIONS

Prior literature suggests many fraud detection techniques using data mining (Fanning *et al.*, 1995, Green *et al.*. 1997, Deshmukh *et al.*, 1997, Fanning *et al.*, 1998, Lin *et al.*, 2003 Bakar *et al.*, 2006). These models require fraud samples (i.e. fraud/none fraud firms), which would make the model inapplicable in other real world settings. This inapplicability stems from the extreme difficulty, if not outright impossibility, of identifying fraudulent firms or transactions with total confidence. Cluster analysis as an unsupervised learning algorithm is a good candidate for fraud and anomaly detection because it sidesteps this difficulty. Our study examines the possibility of using clustering techniques for auditing. Cluster

analysis is applied to a dataset from a major life insurance company in the United States. Claims with similar characteristics are grouped together into clusters. Clusters with small populations and single claims which differ from other claims in the same cluster are flagged for further investigation.

Cluster analysis will always produce grouping. Several parameters are available for the researcher to customize cluster analysis. While one may select different options from others, there is no one correct method. Moreover, the resulting groups may or may not be fruitful for further analysis. Researchers need the expertise of people with domain knowledge for proper evaluation. This study is a preliminary step toward applying cluster analysis in the field of auditing. We show that cluster analysis may be a useful audit technology.

Cluster analysis is a very promising technique that can be integrated into a schema of continuous system monitoring and assurance. Archival studies of data trends will reveal acceptable clusters, problematic ones, and the ability to measure distance from clusters. Experience, judgment, and monitoring procedures will parameterize these clusters, and categorize data. Progressively, clustering findings canl be impounded into *a priori* payment processing filters and block transactions with bad weightings from processing. These filtered transactions will be routed to the continuous auditors (Vasarhelyi, *et al*., 2010 and 2009; Vasarhelyi and Halper, 1991) for review and subsequent action.

7. REFERENCES

BAKAR, Z.; MOHEMAD A, R.; AHMAD A.; DERIS M. M. (2006): "A Comparative Study for Outlier detection Techniques in Data Mining", *Proceeding of IEEE Conference on Cybernetics and Intelligent Systems*.

BROCKETT, P. L.; XIA X.; DERRIG R. A. (1998): "Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud", *Journal of Risk and Insurance*, vol. 65, n.2: 245-274. http://dx.doi.org/10.2307/253535

CHANDOLA, V.; BANERJEE A.; KUMAR V. (2009): "Anomaly Detection: A Survey", *ACM Computing Surveys*, vol. 41, n. 3: 1-58. http://dx.doi.org/10.1145/1541880.1541882

CHAUDHARY, A.; SZALAY A. S.; MOORE A. W. (2002): "Very fast outlier detection in large multidimensional data sets", *Proceeding of ACM SIGMOD Workshop in Research Issues in Data Mining and Knowledge Discovery (DMKD), ACM Press.*

CHEN, M. C.; WANG R. J.; CHEN A. P. (2007): "An Empirical Study for the Detection of Corporate Financial Anomaly Using Outlier Mining Techniques", *Proceeding of the International Conference on Convergence Information Technology*.

CLEARY, B.; THIBODEAU J. C. (2005): "Applying Digital Analysis Using Bedford's Law to Detect Fraud: The Dangers of Type I Errors", *Auditing: A Journal of Practice and Theory*, vol.24, n.1: 77-81.

DAVIDSON, I. (2002): "Visualizing Clustering Results", *Proceeding SIAM International Conference on Data Mining at the University of Illinois*.

DESHMUKH, A.; TALLURU T. (1997): "A Rule Based Fuzzy Reasoning System for Assessing the Risk of Management Fraud", *Journal of Intelligent Systems in Accounting, Finance and Management*, vol.7, n.4: 669-673.

DUAN, L.; XU, L.; LIU Y.; LEE J. (2009): "Cluster-based Outlier detection", *Annals of Operational Research*, vol. 168: 151-168. http://dx.doi.org/10.1007/s10479-008-0371-9

ERTOZ, L.; STEINBACH, M.; KUMAR V. (2003): "Finding Topics in collections of documents: A shared nearest neighbor approach", *Clustering and Information Retrieval*: 83-104.

ESTER, M.; KRIEGEL, H. P.; SANDER J.; XU X. (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proceeding of Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Portland, Oregon: 226-231.

FANNING, K. M.; COGGER, K. O. (1998): "Neural Network Detection of Management Fraud Using Published Financial Data", *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol.7, n.1: 21-41. http://dx.doi.org/10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K

FANNING, K.; COGGER, K. O.; SRIVASTAVA, R. (1995): "Detection of Management Fraud: A Neural Network Approach", *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 4, n. 2: 113-126.

GREEN, B.; CHOI, J. (1997): "Assessing the Risk of Management Fraud through Neural Network Technology", *Auditing*: *A Journal of Practices and Theory,* vol. 16, n.1: 14-28.

GUHA, S.; RASTOGI, R.; SHIM K. (2000): "ROCK, A robust clustering algorithm for categorical attributes", *Information Systems,* vol. 25, n.5, 345-366. http://dx.doi.org/10.1016/S0306-4379(00)00022-3

HAWKINS, D. (1980): "*Identification of Outliers*", Chapman and Hall, London.

HE, Z.; XU, X.; DENG S. (2003): "Discovering cluster-based local outliers", *Pattern Recognition Letters,* vol. 24, n. 9-10: 1641-1650. http://dx.doi.org/10.1016/S0167-8655(03)00003-5

KACHIGAN, S. K. (1991): "*Multivariate Statistical Analysis: a Conceptual Introduction*", Radius Press. New York.

KOHONEN, T. (1997): "*Self-Organizing Maps*", Springer-Verlag New York Inc., Secaucus, New Jersey.

LABIB, K.; VEMURI, R. (2002): "*Nsom: A Real-Time Network-Based Intrusion Detection using Self-Organizing Maps*", Technical Report, Department of Applied Science, University of California, Davis.

LIN, J. W.; HWANG, M. I.; BECKER J. D. (2003): "A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting", *Managerial Auditing Journal* , vol. 18, n.8: 657-665. http://dx.doi.org/10.1108/02686900310495151

RAMADAS, M.; OSTERMANN, S.; JEDEN, B. C. (2003): "Detecting Anomalous Network Traffic with Self-Organizing Maps", *Proceeding of Recent Advances in Intrusion Detection*: 36-54.

ROIGER R. J.; GEATZ M. W. (2003): "*Data Mining: A Tutorial-Based Primer*" (International Edition), Pearson Education, USA.

SHEIK-HOLESLAMI, G.; CHATTERJEE, S.; ZHANG A. (1998): "Wavecluster: A multi-resolution clustering approach for very large spatial databases", *Proceedings of the 24rd International Conference on Very large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco: 428-439.

SMITH, R.; BIVENS, A.; EMBRECHTS, M.; PALAGIRI, C.; Szymanski B. (2002): "Clustering approaches for anomaly based intrusion detection", *Proceedings of Intelligent Engineering Systems through Artificial Neural Networks*, ASME Press: 579-584.

SUN, H.; BAO, Y.; ZHAO, F.; YU, G.; WANG, D. (2004): "Cd-trees: An efficient index structure for outlier detection", *Lecture Notes in Computer Science*, vol. 3129: 600-609. http://dx.doi.org/10.1007/978-3-540-27772-9_60

TANG, P-N: STEINBACH, M.; KUMAR, V. (2006): *"Introduction to Data Mining"*, Pearson Education, Inc.

VASARHELYI, M.; WARREN Jr., J.; TEETER, R.; TITERA, W. (2011): "Embracing the Automated audit: how common data and audit apps will enhance auditor judgment and assurance", Working paper, CarLab, Rutgers Business School. http://raw.rutgers.edu

VASARHELYI, M; ALLES, M.; WILLIAMS, K.T. (2010): "Continuous assurance for the now economy", *A Thought Leadership Paper for the Institute of Chartered Accountants in Australia*, Melbourne. Accessible at: http://bit.ly/ocbx2l

VASARHELYI, M.; KUENKAIKAEW, S.; ROMERO, S. (2009): "Continuous Auditing and Continuous Control Monitoring: Case studies from leading organizations", Working Paper, Rutgers Accounting Research Center. Accessible at: http://bit.ly/oYiHyT

VASARHELYI, M.; HALPER, F. (1991): "The Continuous Audit of Online Systems", *Auditing: A Journal of Practice and Theory*, vol. 10, n.1: 110-125.

YU, D.; SHEIKHOLESLAMI, G.; ZHANG, A. (2002): "Findout: finding outliers in very large datasets", *Knowledge and Information Systems*, vol.4, n.4: 387-412. http://dx.doi.org/10.1007/s101150200013