Cluster Analysis for Anomaly Detection in Accounting Data

Sutapat Thiprungsri, Rutgers University, Newark, NJ, USA. (sutapat@pegasus.rutgers.edu)

Abstract: Cluster Analysis is a useful technique for grouping data points such that points within a single group or cluster are similar, while points in different groups are distinctive. Clustering as an unsupervised learning algorithm is a good candidate for fraud and anomaly detection. The purpose of this study is to examine the possibility of using clustering technology for continuous auditing. Automating fraud filtering can be of great value to preventive continuous audits. In this paper, cluster-based outliers help auditors focus their efforts when evaluating group life insurance claims. Claims with similar characteristics have been grouped together and those clusters with small population have been flagged for further investigations. Some dominant characteristics of those clusters are, for example, having large beneficiary payment, having huge interest amount and having been submitted long time before getting paid.

This study examines the application of cluster analysis in accounting domain. The results provide a guideline and evidence for the potential application of this technique in the field of audit.

I. Introduction

Clustering is an unsupervised learning algorithm, which means that there is no label (class) for the data (Kachigan, 1991). Clustering is a useful technique for grouping data points such that points within a single group or cluster are similar, while points in different groups are dissimilar. In general, the greater similarity within a group and the greater differences between groups mean the better clustering results.

There is no absolute best clustering technique. Users' needs are an important factor in evaluating the clustering technique. The best techniques provide the results that are useful for the user's purposes. Moreover, the cluster evaluation is quite subjective because the results can be interpreted in different ways. Several factors should be considered when deciding upon which type of clustering technique to use. These factors include type of clustering techniques, characteristics of clusters, characteristics of the data set and attributes, noise and outliers, the number of data objects, the number of attributes, cluster description and algorithm consideration (Tang et al, 2006).

Clustering as an unsupervised learning algorithm is a good candidate for fraud and anomaly detection techniques because it is difficult to identify suspicious transactions. Clustering could be used to group transactions so that different attention and effort could be applied to each different cluster.

The purpose of this study is to examine the possibility of using clustering technique for continuous auditing. I apply cluster analysis to a unique dataset provided by a major insurance company in the United States and examine the cluster-based outliers.

Group life insurance claims have been grouped into clusters of claims, and claims with similar characteristics have been grouped together. Those clusters with small populations have been flagged for further investigations.

II.Literature Review

Fraud Detection

In the accounting literature, most studies focus on management fraud. For the prediction of management fraud, most prediction models employ either logistic regression or the Neural Network as the technique.

Bell et al (2000) present the results of an attempt to develop a model useful in predicting the existence of fraudulent financial reporting. The author proposes a working discriminant function for the conceptual model from Loebbecke et al (1989). Using a sample of 77 frauds, and 305 non-fraud control firms, Bell et al (2000) develop and test logistic regression model that estimate the likelihood of fraudulent financial reporting for an audit client, conditioned on the presence of fraud-risk factors.

Fanning et al (1995) propose an alternative approach, Artificial Neural Networks (ANNs), for detection of management fraud. Neural networks are designed using both generalized adaptive neural network architectures (GANNA) and the Adaptive Logic Network (ALN). Using the same data set as Bell et al (2000), the prediction accuracy is 89% for GANNA and 90% from ALN.

Green et al. (1997) examine the use of neural networks (NN) as a means of detecting financial statement fraud in the revenue and collection cycle of publicly held manufacturing and merchandising companies. Five ratio (Allowance for doubtful account/ Net sales, Allowance for doubtful account/ AR, Net sales/ AR, Gross Margin/ Net sales, AR/ TA) and three accounts (Net sales, AR, Allowance for doubtful account) are used. Eighty-six (86) fraud firms and 86 non-fraud firms are used as samples. The different models' performance, or accuracy, ranks from 32% to 62%.

Deshmukh et al. (1997) develop membership functions and fuzzy rules for assessing risk of management fraud using the statistical significance of each red flag and theoretical model. Using the same data set as Bell et al (2000), the model produces a similar result.

Fanning et al. (1998) propose the use of self-organizing Artificial Neural Network (ANN), AutoNet, to develop a model for detecting management fraud. The paper applies the technique to the publicly available financial information. From twenty variables that are possible indicators of fraudulent financial statement, the neural network model selected a discriminant function that was statistically successful on a holdout sample. The model's prediction accuracy was 63%. The neural net model performs better than linear and quadratic discriminant analysis and logistic regression.

Lin et al (2003) evaluate the utility of an integrated fuzzy neural network (FNN) for fraud detection. FNNs are a class of hybrid intelligent systems that integrate fuzzy logic with Artificial Neural Network. The variables used are all financial ratios. The FNN developed in this research outperformed most statistical models and artificial neural networks (ANNs) with approximately 76% accuracy.

Though several fraud prediction models mentioned previously provide very good prediction performance, the majority of papers extend the work of Bell et al. (1991). Three of the papers (Bell et al, 2000, Fanning et al, 1995, Deshmukh et al., 1997) use the same data set to test for model performances. The data used is a set of questions answered by partners of KPMG Peat Marwick. These models have two major disadvantages derived from the use of this data set. First, these three models' performances could have been overstated due to a possible hindsight bias inherent in the judgment made by the auditors associated with fraud engagements (Bell et al, 2000). Finally, the data is not publicly available. The researchers have to spend time to collect. Therefore, the application of these models to general cases may be difficult, if not impossible, and it might not be cost effective to do.

The prediction rates of all other models generally range from 50-65%. This level of prediction performance is not high. Although Lin et al (2003) show a generally high prediction performance, the use of only groups of variables related to accounts receivable is a limitation. Moreover, the number of observations in the fraud sample used in the paper is small. Accounts receivable has already been proven as a good predictor of fraud. Using variation of AR-related ratio would not contribute greatly to the literature. Better, more accurate models for fraud predictions are desired. Moreover, most models aim at predicting management frauds.

Cluster Analysis

Cluster analysis groups the objects based only on information found in the data that describes the objects and their relationships (Tan et al, 2006). The greater similarity within a group and the greater differences between groups mean the better clustering results. It begins with an undifferentiated or single group, followed by attempt to form subgroups, which are differentiated by the selected variables.

Clustering can be used for data exploration and also to understand the structure of data. It is used to find similarities between observations and then group them. The two major steps in cluster analysis are 1) selecting measures of similarities or dissimilarities, and 2) selecting the procedures for cluster formations (Kachigan, 1991). There are several options or techniques available for these steps, making cluster analysis as much an art as a science. The purpose of performing cluster analysis is to ask the question whether a given group can be partitioned into different subgroups. The subgroups or clusters can be named or defined using the group mean as the representative of the observations in the group.

Clustering is a widely used technique in the area of marketing research especially market segmentation, market structure analysis and a study of customer behaviors. For example, in marketing customer segments (or clusters) are defined using demographic information. Different marketing strategies are then developed and applied to each customer segments or clusters. In the review of cluster analysis in marketing research, Punj et al. (1983) list many marketing literatures prior to 1983 that applied cluster analysis as their methodologies for understanding market segments and buyer behaviors. Market segmentations using cluster analysis have been examined in many different industries, including finance and banking (Anderson et al, 1976, Calantone et al, 1978), automotive (Kiel et al, 1981), education (Moriarty et al, 1978), consumer products (Sexton, 1974, Schaninger et al, 1980) and high technology industry (Green et al, 1968).

Steps in Cluster Analysis

Cluster analysis is generally started with observation measurements. Observations are measured on K variables. The observations' similarities, or distances between each pair of observations, are measured. An algorithm will be employed to group the observations into subgroups based on those observations similarities. Then clusters are formed. The goal is to create clusters that have small within--cluster variation, but large between-cluster variation. Finally, clusters are compared. The differences between clusters can be seen from their representative values such as mean values of the input variables from each cluster. The steps are shown in Figure 1.

Unlike other data analysis methods, cluster analysis has been developed throughout the years in many disciplines. There is no single dominant discipline. The purpose and benefit of the cluster analysis depend on the type of the applications. For examples, in marketing, cluster analysis may be used mainly to learn about market segments and to seek a better understanding of buyer behaviors by identifying homogeneous groups of buyers in many different industries (Punj et al, 1983).

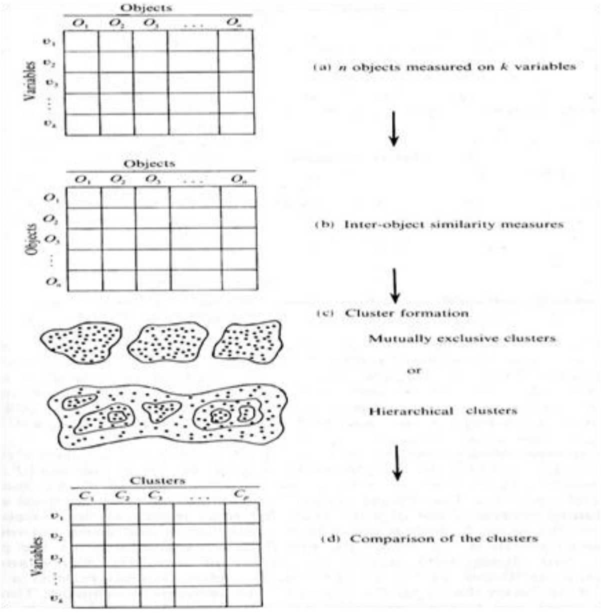

Objects
O_1 O_2 O_3 . . . O_n
Variables
v_1 v_2 v_3 ⋮ v_k
(a) n objects measured on k variables
Objects
O_1 O_2 O_3 . . . O_n
Objects
O_1 O_2 O_3 ⋮ O_n
(b) Inter-object similarity measures
(c) Cluster formation
Mutually exclusive clusters
or
Hierarchical clusters
Clusters
C_1 C_2 C_3 . . . C_p
Variables
v_1 v_2 v_3 ⋮ v_k
(d) Comparison of the clusters

Figure 1. An Outline of Cluster Analysis Procedure. (Kachigan, 1991).

III. The Setting: Group Claims

This study evaluates group life claims offered by a major insurance company to employees of client companies. Group life is the type of group insurance that the insurance company markets to corporations. The raw dataset for this study is from a group life claims business unit of a major insurance company in the United States.

Group life insurance is different from individual life insurance in many ways. For example, group life insurance is sold to companies in volume i.e. company A buys group life insurance for 100 employees; while individual life insurance is sold to an individual e.g. Mr. B buys life insurance for himself. From the perspective of the insurance provider, the purchasing company is the customer in the former case, while the individual is the customer in the latter case. Company A offers the insurance to its employees as a work benefit. The employees might also have the option to purchase an additional individual life insurance policy. Because of the differences mentioned, the insurance company manages policies and claims from these two types of life insurance differently. While the insurance company collects an insured person’s information for individual life insurance policies, they do not keep that level of information for group life insurance policies. Rather, the information about a specific employee (or the insured) is entered into the system only once a claim is received. If company A submits a claim, the insurance provider would take the information from the company as is. There is very little verification done in the part of the provider. For example, if company A submits a life claim for Mr. John, the insurance provider would check the employee’s death against the Social Security Administration’s death file (SSA death file). There is no further verification on the real existence of that employee in the company.

The nature of the group life insurance can bring about many risks into the policy administration and the audit. Therefore, the insurance company is seeking for an innovative ways which would help to control and reduce the risk of fraudulent claims. The purpose of this study is to apply the use of cluster technology to provide the aid for the internal auditor in the internal audit process.

Data

The data set contains the records of group life claim payments which were paid out in 2009. The data contains 208 attributes related to group life claims, and these attributes are mainly composed of 5 groups.

Attributes related to the insured

Attributes related to the coverage

Attributes related to the group / company

Attributes related to the beneficiary

Attributes related to the payment

The company receives data in paper form. Clients submit the paper claim document to the insurance company, and the insurance company scans the claim document into the system as a PDF file. The information is subsequently input into the BIOS systems manually. Because the data is manually input into the system, several mistakes are found in the data, including wrong dates, typing errors, and misspellings. In some cases, the wrong information can be easily identified; for example, when default dates have been used for insured birth dates, death dates, and/or hiring dates. If the birth dates entered are only a few days or a few months before the death date or hiring dates, they are clearly invalid data. The reason is that, according to the law, companies can not hire anyone who is younger than a certain age (i.e. an infant cannot be an employee in any company). The invalid information indicates that the system has the data quality issue.

Each claim received is identified by a claim id (CLM_ID). However, in the BIOS systems, each record represents individual payments. These payments could be beneficiary payment or interest. A claim could have multiple beneficiaries and/or multiple coverage policies. It is normal that more than one record would be related to a single CLM_ID. Therefore, in the original data set, the CLM_ID would not be a unique identifier of the record. Because a claim received is processed as one, regardless of the number of related beneficiaries or coverage policies, the analysis is performed based on individual claim. Each claim is approved, denied or frozen by CLM_ID. In other words, each claim received is evaluated by its CLM_ID.

The attributes in the data also reveal inconsistencies. From the original set of the attributes, many attributes are left blank. The possible reasons are

(1) these are real missing values and (2) the attributes are not used in each specific claim. From the total 208 attributes, only 65 attributes have fewer than 15% missing values and have less variety of unique values (less than 500). From these 65 attributes, five (5) attributes lack variety in their values e.g. over 99% of the records have the same values. Most attributes contain specific information such as name and address of insured, beneficiary and dependent.

Clustering Procedure

The sample contains 40,080 group life insurance claims paid out in 2009. Because a claim can have multiple beneficiaries and/or expanded coverage, there are over 65,000 payments related to these 40,080 claims. No matter how large the coverage and number beneficiaries are, claims are evaluated as one. Therefore, the analysis will be done claim by claim.

The software used for this analysis includes SAS and Weka. The data set was cleaned and transformed using SAS. The clean data was then exported into a comma separated value (CSV) file. Then the dataset was prepared in the ARFF format in order to be fed into Weka

Because of data quality issues mentioned in the previous section and the fact the extended coverage and/or multiple beneficiary claims are evaluated as one, a new dataset was created based on the original data. For the first step, four (4) newly created attributes are selected as the attributes for clustering. They are as following

- Percentage: Total interest payment / Total beneficiary payment
- AverageCLM_PMT: Average number of days between the claims received date and the payment date (a weighted average is used because a claim could have multiple payment dates)
- DTH_CLM: Number of days between the death date and the claim received date.
- AverageDTH_PMT: Average number of days between the death date and the payment date (a weighted average is used because a claim could have multiple payment dates)

These attributes are on different scales. Therefore, they were normalized so they could be compared. The differences in scale would then have less affect on the results. Because all five attributes are numeric, simple K-mean clustering is selected for the clustering procedure. K-mean clustering is a very simple yet well known algorithm for clustering. It is much less computer intensive than many other algorithm, therefore, it is usually a preferable choice when the dataset is large. Because K-mean does not have the option to automatically use the number of clusters to group the data, the number of cluster must be pre-selected. Therefore, the number of clusters must be selected first. No matter how many clusters the dataset is grouped to, the following steps are generally the same. The second step for K-mean clustering is initially put each observation into clusters. This step can be done randomly or systematically. A centriod of clusters is selected, and each observation is assigned to the closet cluster. The third step computes the distance between each observation and the cluster centers. If the observation is not currently a member of the cluster with the closet centroid, the observation is reassigned to a new cluster. The former cluster loses membership, while the new cluster with the closet centriod gains membership. The centroids are then recalculated. The process from step three repeats until there is no new assignment. The algorithm for K-mean clustering is explained 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.*
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 center and repeat steps 3-5.*

For this sample, several numbers of clusters have been tested. Two different combinations of attributes are used for clustering. First, the Percentage and AverageDTH_PMT are used for the first clustering. The number of clusters which create the changing point for the mean squared error is selected. For this combination of attributes, the number of clusters selected is eight (8). Second, all four variables are used for clustering. The number of clusters selected is thirteen (13).

Anomaly Detection

According to American Heritage College dictionary (2004) error and anomaly are defined as follows:

Error n. 1. An act, assertion, or belief that unintentionally deviates from what is correct, right or true. 2. The conditional of having incorrect or false knowledge. 3. The act of an instance of deviating from an accepted code of behavior. 4. A mistake.

Anomaly n. 1. Deviation or departure from the usual or common order, form or rule. 2. One that is peculiar, irregular, abnormal, or difficult to classify.

The distinguishing factor between fraud and error is intention. Error is the deviation from usual behavior. It does not have the element of intention to deviate. Therefore, a deviation or anomaly could be a result of an error or the intention to commit fraud.

Anomalies occur for many reasons. For example, data may come from different classes, natural variation in the data and data measurement, or collection error (Tang et al, 2006).

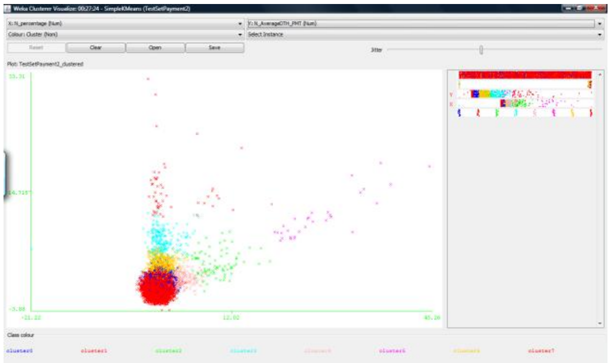
Outliers are observations that deviate so much from other observations that they arouse suspicion that they were generated by a different mechanism (Hawkins, 1980). They are traditionally considered to be single points. Duan et al (2009) suggests that there is a possibility that many abnormal events have both temporal and spatial locality, which might form small clusters that also need to be deemed as outliers. In other words, not only single points but also small clusters can probably be considered outliers. This type of outlier is called a "cluster-based outlier".

This paper examines both individual observations and small clusters as possible outliers. Most data points in the dataset should not be outliers. The outliers are then identified in two ways. First, observations that have low probability of being a member of a cluster (i.e. are far away from other members of the clusters) are identified as outliers. The probability of 0.6 is used as a cut-off point. Second, the clusters that have small populations should possibly be considered outliers. In this aspect, clusters populated with less than 1% of the whole population are considered as outliers.

Result

Because of the simplicity and the 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 results from Weka are represented in Table 1 and Table 2.

For the first set of clusters using two (2) attributes, eight (8) clusters are formed. About 90% of claims are clustered into cluster 7 and 6% are in cluster 0 (from Table 1). Three clusters (1, 2, and 5) have membership of less than 1%. The numbers of claims in those clusters are 54, 84 and 31 respectively. Examining the characteristics of these less populated clusters, a couple suspicious characteristics should be mentioned. Claims in these clusters have high interest/beneficiary payment percentage and/or claims with a 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 2.


Weka Clusterer Visualize: 00:27:24 - SimpleKMeans (TestSetPayment2)

X: N_percentage (Num)

Y: N_AverageDTH_PMT (Num)

Colour: Cluster (Nom)

Plot: TestSetPayment2_clustered

Figure 2. Visualization of the Cluster Assignment for 2 attributes clustering; N_Percentage and N_AverageDTH_PMT.

For the second set of clusters using four (4) attributes, thirteen (13) clusters are formed. About 81% of claims group into cluster 8. Six clusters are populated with less than 1% of the claims. These are clusters 2, 3, 5, 10, 11 and 12. The numbers of claims in those clusters are 194, 98, 30, 39, 110 and 97, respectively. Because of time and budget constraints, it is not wise to follow up on all of the claims in the suspicious clusters. Therefore, not all small clusters are selected for further investigation. Moreover, some suspicious claims may have valid reasons. For examples, claims that have been the system for a long time before the beneficiaries get paid may seem suspicious; however, a possible explanation might be that the paperwork was not completed. All the small clusters should be closely examined. From these six small clusters, the one that should be selected for the follow up is cluster 12. This cluster has a high interest/beneficiary percentage, while the length of time from the death to payment date is not as high. These should raise questions concerning the reason for the high interest. In addition to identifying small clusters as possible anomalies, the probability of individual observations' cluster membership is examined. Using probability of 0.6 as the cutoff point, 547 claims are identified as possible anomalies. Clusters with larger membership have higher numbers of possible anomalies. The visualized results are shown in Figure 3.

Table 1: Result of Cluster Analysis using two attributes from Weka

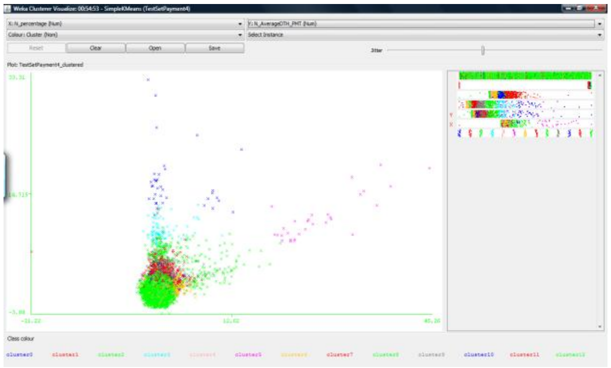
```
=== Run information ===
Scheme:      weka.clusterers.SimpleKMeans -N 8 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:    TestSetPayment2
Instances:   40080
Attributes:  3
             N_AverageDTH_PMT
             N_percentage
Ignored:
             CLM_ID
Test mode:   evaluate 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#
                                        Full
Attribute                               Data        0         1         2         3         4         5         6         7
                                     (40080)   (2523)      (54)      (84)     (222)     (295)      (31)     (768)   (36103)
----------------------------------------------------------------------------------------------------------------------------
N_AverageDTH_PMT                      0.0004   0.6374    15.177    3.5419    6.9858    0.8778   10.9006    2.7806   -0.1937
N_percentage                         -0.0013   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%)
```


Weka Clusterer Visualize: 00:54:53 - SimpleKMeans (TestSetPayment4)

X: N_percentage (Num)

Y: N_AverageDTH_PMT (Num)

Colour: Cluster (Nom)

Plot: TestSetPayment4_clustered

Figure 3. Visualization of the Cluster Assignment for 4 attributes clustering: N_Percentage, N_AverageDTH_PMT, N_AverageCLM_PMT, N_DTH_CLM.

In order to verify if the cluster analysis could really identify anomaly in the accounting system, the suspicious cluster/individual claims should be selected for the further investigation by the internal auditor. The results from the follow up would help to improve the model.

IV. Conclusion

Because it is difficult to identify suspicious transactions, cluster analysis as an unsupervised learning algorithm is a good candidate for fraud and anomaly detection techniques. Clustering could be used to group transactions so that different attention and efforts could be applied to each different cluster. This study examines the possibility of using clustering technique for continuous auditing. Cluster analysis is applied to a data set from a major life insurance company in the United States. Cluster-based outliers were examined. Group life insurance claims were grouped into clusters of claims. Claims with similar characteristics were grouped together. Clusters with small populations were flagged for further investigations.

Cluster analysis will always produce grouping. Several options and/or parameters are available for researcher to choose in order to perform cluster analysis. One may select different options from others. It does not always mean that one is right and the other is wrong. Moreover, the resulting groups may or may not be meaningful for further analysis. To clearly evaluate the results, researchers need helps from people with domain knowledge.

This study is a preliminary step to apply the cluster analysis in the field of continuous auditing. It shows that cluster analysis may be useful technology for accounting.

Table 2: Result of Cluster Analysis Using 4 attributes from Weka

```
=== Run information ===
Scheme:     weka.clusterers.SimpleKMeans -N 13 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:
TestSetPayment4
Instances:    40080
Attributes:   5
            N_AverageCLM_PMT
            N_DTH_CLM
            N_AverageDTH_PMT
            N_percentage
Ignored:
            CLM_ID
Test mode:    evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 110
Within cluster sum of squared errors:
8.938107429242356
Missing values globally replaced with mean/mode
Cluster centroids:
                Cluster#
                        Full
Attribute       Full    Data       0      1      2      3      4      5      6      7       8      9     10     11     12
                     (40080)   (510)  (343)  (194)   (98) (3699)   (30) (1275)  (741) (32658)  (286)   (39)  (110)   (97)

N_AverageCLM_PMT           0    3.33   5.85   1.12   0.93   0.27   1.08   1.44  -0.02   -0.26   0.33   1.28   9.81   4.04
N_DTH_CLM                  0    0.05   0.29   5.63   9.27  -0.10  11.51  -0.11   0.83   -0.13   2.89  17.31   0.40   0.49
N_AverageDTH_PMT           0    1.24   2.37   5.64   8.93   0.01  11.06   0.40   0.78   -0.21   2.79  16.50   3.80   1.90
N_percentage               0    0.21   0.16   1.78   0.66   0.11  26.89   0.51   0.48   -0.12   1.00   2.22   0.30   7.78

Clustered Instances
 0      510 (  1%)
 1      343 (  1%)
 2      194 (  0%)
 3       98 (  0%)
 4     3699 (  9%)
 5       30 (  0%)
 6     1275 (  3%)
 7      741 (  2%)
 8    32658 ( 81%)
 9      286 (  1%)
10       39 (  0%)
11      110 (  0%)
12       97 (  0%)
```

Table 3. Results

Cluster Analysis	Cluster-Based Outliers	Distance-Based Outliers
Cluster Analysis with 2 Attributes	169	325
Cluster Analysis with 4 Attributes	568	547

V. References

American Heritage College Dictionary, 2004, USA, p.58, p.475

Anderson, W. T. Jr., E. P. Cox, III and D. G. Fulcher, 1976, Bank Selection Decisions and Market Segmentation, Journal of Marketing, 40, p.40-45

Bell, T.B., Szykowny S. and Willingham, J.J., 1991, Assessing the Likelihood of Fraudulent Financial Reporting: A Cascaded Logit Approach. Working paper, KPMG Peat Marwick. Montvale. NJ.

Bell, T.B. and J.V., Carcello, 2000, A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting, Auditing: A Journal of Practice and Theory 19(1): 169-184

Calantone, R. J. and A. G. Sawyer, 1978, Stability of Benefit Segments, Journal of Marketing Research 15(August), p.395-404.

Deshmukh, A. and T. Talluru, 1997, A Rule Based Fuzzy Reasoning System for Assessing the Risk of Management Fraud, Journal of Intelligent Systems in Accounting, Finance & Management 7(4): 669-673.

Duan, L., L. Xu, Y. Liu and J. Lee, 2009, Cluster-based Outlier detection, Annals of Operational Research 168: 151-168.

Fanning, K., K. O. Cogger, K. O. and R. Srivastava, 1995, Detection of Management Fraud: A Neural Network Approach, International Journal of Intelligent Systems in Accounting, Finance & Management, 4(2): 113-126.

Fanning, K. M. and K. O. Cogger, 1998, Neural Network Detection of Management Fraud Using Published Financial Data, International Journal of Intelligent Systems in Accounting, Finance & Management 7: 21-41

Green, B. and J Choi, 1997, Assessing the Risk of Management Fraud through Neural Network Technology, Auditing: A Journal of Practices & Theory 16(1): 14-28

Green, P. E. and F. J. Carmone, 1968, The Performance Structure of the Computer Market: A Multivariate Approach, Economic and Business Bulletin, 20, p.1-11.

Hawkins, D., 1980, Identification of Outliers, London: Chapman and Hall.

Kachigan, S. K. 1991, Multivariate Statistical Analysis A Conceptual Introduction, Radius Press

Kiel, G. C. and R. A. Layton, 1981, Dimensions of Consumer Information Seeking Behavior, Journal of Marketing Research, 18(May), p.233-9.

Lin, J. W., M. I. Hwang, and J. D., Becker, 2003, A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting, Managerial Auditing Journal 18(8): 657-665.

Loebbecke, J.K., M.M. Eining, and J.J. Willingham, 1989, Auditors' Experience with Material Irregularities: Frequency, Nature, and Detectability, Auditing: A Journal of Practice & Theory 9 (1): 1-28

Moriarty, M. and M. Venkatesan, 1978, Concept Evaluation and Market Segmentation, Journal of Marketing, 42, p.82-86.

Punj, G. and D. W. Stewart. 1983, Cluster Analysis in Marketing Research: Review and Suggestions for Application, Journal of Marketing Research (May 1983), 134-48.

Roiger, R. J. and M. W. Geatz. 2003, Data Mining A Tutorial Based Primer (International Edition), Addison Wesley.

Schaninger, C. M., V. P. Lessig and D. B. Panton. 1980, The Complementary Use of Multivariate Procedures to Investigate Nonlinear and Interactive Relationships Between Personality and Product Usage, Journal of Marketing Research 17(February), 119-24.

Sexton, D. E. Jr., 1974, A Cluster Analytic Approach to Market Response Functions, Journal of Marketing Research, 11(February), p.109-114.

Tang, P-N, M. Steinbach and V. Kumar. 2006, Introduction to Data Mining, Pearson Education, Inc.