International Journal on Emerging Technologies (Special Issue on NCRIET-2015) **6**(2): 358-363(2015)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Finding anomalies using cluster analysis by cosine similarity measures

Gayatri Mugli
Asst. Prof BKIT Bhalki, INDIA

(Corresponding author: Gayatri Mugli)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In data mining, anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset**. Anomaly detection refers to detecting patterns in a given data set that do not conform to an established normal behavior. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. 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. Anomalies are also referred to as outliers, change, deviation, surprise, aberrant, peculiarity, intrusion, etc. In particular in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.**

Keywords*: Anomaly, Cluster, Similarity, Pattern, Data set.*

I. INTRODUCTION

Data mining : Data mining (the advanced analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection. Anomaly detection is an important problem that has been researched within diverse research areas and application domains for each category we have identified key assumptions, which are used by the techniques to differentiate between normal and anomalous behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. For each category, we provide a basic anomaly detection technique, and then show how the different existing techniques in that category are variants of the basic technique

Anomaly and Anomaly Detection:

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201
424	Dr. Newsome	29-Mar-2007	?

Fig 1: Anomalous data in a data set

Outliers are observations that deviate so much from other observations that they may have been generated by a different mechanism. 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.

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.

Collective anomaly occurs when a collection of related data instances is anomalous. 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.

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 .

A variety of methods for each step has been used in many fields 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.

Fig 2: Graph shows anomaly

This study examines the application of cluster analysis in the data set. In particular its application to discrepancy detection in the field of data set . Clustering is an unsupervised learning algorithm, which means that data are analyzed without the presence of predetermined labels (e.g. "fraudulent/non - fraudulent"). 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 data set. Automated fraud filtering can be of great value as a preventive tool. We apply cluster analysis to a dataset provided by a company and examine the resulting outliers. Cluster-based outliers help auditors focus their efforts when evaluating attribute values. Some dominant characteristics of outlier clusters are given all attribute values maximum range or out of range values, values may be continous or discrete or fraction or in range it is not with respect to attribute type.

II. LITERATURE SURVEY

[1] Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis

This paper mainly focuses on intrusion detection based on data mining. The aim is to improve the detection rate and decrease the false alarm rate, and the main research method is clustering analysis. The algorithm and model of Intrusion Detection are proposed and corresponding simulation experiments are presented. Firstly, a method to reduce the noise and isolated points on the data set was advanced. By dividing and merging clusters and using the density radius of super sphere, an algorithm to calculate the number of the Cluster Centroid was given. By the more accurate method of finding *k* clustering center, an anomaly detection model was presented to get better detection effect. This paper used KDD CUP 1999 data set to test the performance of the model. The results show the system has a higher detection rate and a lower false alarm rate, it achieves expectant aim.

[2] Outlier Detection in Data Streams Using Various Clustering Approaches

In the case of data streams, the goal of outlier detection algorithm is to detect the outlier up to N number of data chunks. Numerous methods have been proposed till today. Several of the clustering methods produce a set of methodology. But still it is difficult to recommend any one technique as superior, that also depend on the choice of dataset being used. So, there is no single outlier detection algorithm that is best of all kind of dataset, because every algorithm has there pros and cons. If we talk about clustering in arbitrary shapes then density based is the best. If particular dataset is considered and it is complex or large then partitioning based cluster has higher computational cost that is the disadvantage for it and hierarchical is slow for large dataset. So, this paper provides the concept behind data streams, different Clustering techniques and difference between them.

[3]A Comparative Study for Outlier Detection Techniques in Data Mining

This paper presented the result of an experimental study of some common outlier detection techniques. Firstly, we compare the two outlier detection techniques in statistical approach, linear regression and control chart techniques. The experimental results indicate that the control chart technique is better than that liner regression technique for outlier data detection. Next, we analyze Manhattan distance technique based on distance-based approach. The experimental studies shows that Manhattan distance technique outperformed the other techniques (distance-based and statistical-based approaches) when the threshold values increased

III. OBJECTIVE

A specific(result) anomalies that a person or system aims to achieve within a time frame and with available resources.

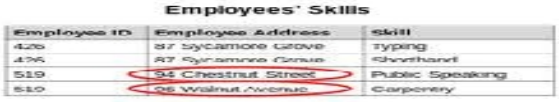

Fig 3: Anomalous data in a data set

Forecasting what may happen in the future. Classifying people or things into groups by recognizing patterns. Clustering people or things into groups based on their attributes. Sequencing what events are likely to lead to later events. . In data set, get optimal solution for anomalies.

Particular in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object.

IV. METHODOLOGY

A system of methods used in a particular area of study or activity. Methodology is the systematic, theoretical analysis of the methods applied to a field - for example, a field of study, or the field of how to calculate some particular value. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. It, typically, encompasses concepts such as paradigm, theoretical model, phases and quantitative or qualitative techniques.

A methodology does not set out to provide solutions - it is, therefore, not the same thing as a method. Instead, it offers the theoretical underpinning for understanding which method, set of methods or so called "best practices" can be applied to a specific case, for example, to calculate a specific result.

Sample Anomaly Detection Problems

These examples show how anomaly detection might be used to find outliers in the training data or to score new, single-class data.

Figure 1 Sample Build Data for Anomaly Detection

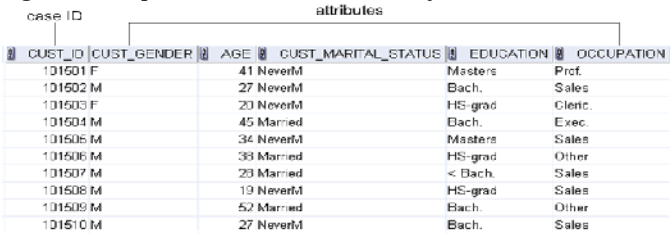

Fig 4: data for anomaly detection

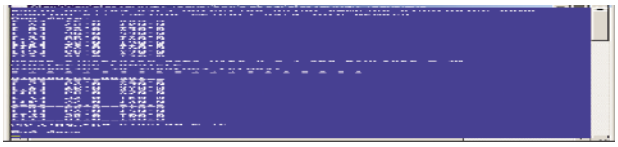

Fig 5: Anomalous data in a data set

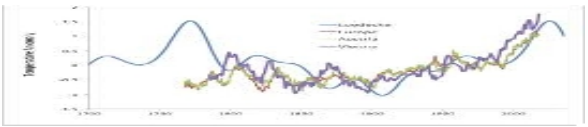

Fig 6: Graph for Anomalous data

Clustering based techniques for anomaly detection:

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, ROCK-Robust Clustering using links , SNN cluster-Shared Nearest Neighbor Clustering ,Find Out algorithm and Wave Cluster algorithm . 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*. Self-Organizing Maps (SOM) are used for anomaly detection in many different applications, including fraud detection) and network intrusion. The techniques in this group involve two steps: grouping data into clusters, and calculating distances from cluster *centroid*s to identify anomaly scores. 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. A technique called Find CBLOF 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. 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. A technique called CD-trees. Both techniques define sparse clusters as anomalies.

Similarity measures for cluster analysis:

Similarity measure: In computer science, a similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. Although no single definition of a similarity measure exists, usually similarity measures are in some sense the inverse of distance metrics: they take on large values for similar objects and either zero or a negative value for very dissimilar objects. E.g., in the context of cluster analysis, Frey and Dueck suggest defining a similarity measure

$$s(x_i, x_k) = -||x_i - x_k||_2^2$$

where $||x_i - x_k||_2^2$ is the squared Euclidean distance.

In information retrieval, cosine similarity is a commonly used similarity measure, defined on vectors arising from the bag of words model. In machine learning, common kernel functions such as the RBF kernel can be viewed as similarity functions.

Cosine similarity measures: it is used for document similarity .if x and y are two document vectors,then

$$\cos(x, y) = \frac{x.y}{||x||\ ||y||}$$

Where . (dot) indicates the vector dot product ,

$$x.y = \sum_{k=1}^{n} x_k\ \ y_k$$

x_k y_k, and IIx II is the length of vector x,

$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x.x}$$

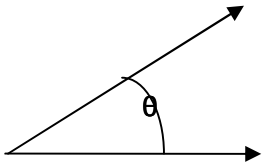

Fig 7: Geometric illustration of the cosine measure

Above fig indicates, cosine similarity really is a measure of the (cosine of the) angle between x and y . Thus, if the cosine similarity is 1 , the angle between x and y is 0 degree ,and x and y are the same except for magnitude (length). If the cosine similarity is 0 ,then the angle between x and y is 90 degree, and they not share any terms(words).

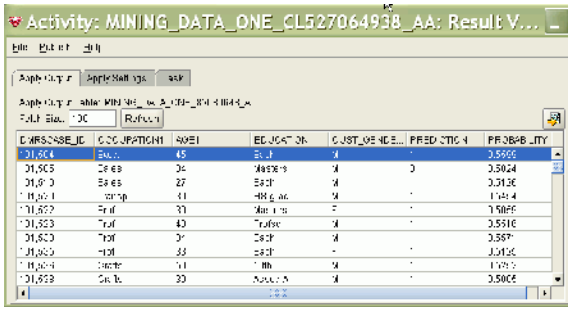

Fig 8: Outliers in the data

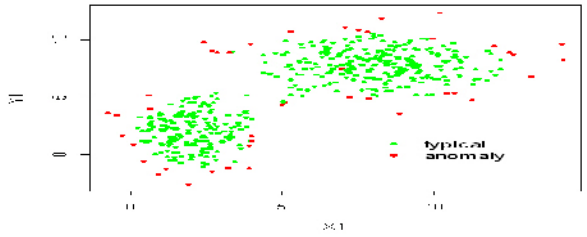

Fig 9: Graph of anomaly

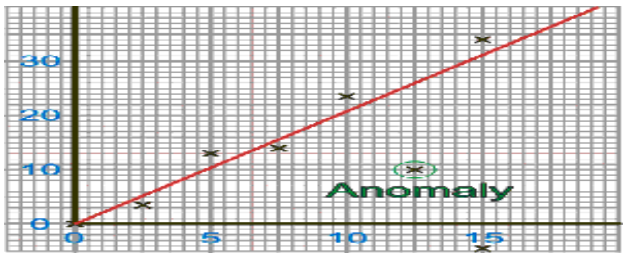

Fig 10: Graph of anomaly

Above equation can be written as

$$\cos(x, y) = \frac{x.y}{||x||\ ||y||} = x' \,.\, y'$$

Where x' = x/ ||x|| and y' = y/ ||y|| .Dividing x and y by their lengths normalizes them to have a length of 1. This means that cosine similarity does not take the magnitude of the two data objects into account when computing similarty. (Euclidean distance might be a better choice when magnitude is important.)For a factors with a length of 1, the cosine measure can be calculated by taking a simple dot product .Consequently ,when many cosine similarities between objects are being computed ,normalizing the objects to have unit length can reduce th time required.

POSSIBLE OUTCOME

This paper proposed the cosine similarity method to know that the similar data set are in clusters and anomalous are having in a cluster of smaller size. As for the future work, we plan on extending current work to accommodate different similarity measures to find anomalies, by using different size data set by using different cluster techniques and by different similarity measures for finding anomalies.

6. CONCLUSIONS

From the above paper we come to know that if we are applying proper method than we can get optimal solution in finding the anomalies using cluster analysis by cosine similarity measures. 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.

REFERENCES

[1] Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis ,LI Han School of Science, Beijing Information Science and Technology University Beijing, P.R.China,

[2] Outlier Detection in Data Streams Using Various Clustering Approaches Kusum Makkar ,Meghna Sharma Assistant Professor, ITM University Gurgaon, INDIA, ieee communication surveys & tutorials, vol. 17, no. 1, first quarter 2015

[3] 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*.

[4] 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 82 The International Journal of Digital Accounting Research Vol. 11

[5] 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.*

[6] 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*.

[7] 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.

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

[9] 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.

[10] 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

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

[12] 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.

[13] 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

Thiprungsri & Vasarhelyi Cluster Analysis for Anomaly Detection… 83.

[14] 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.

[15] 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