FRAUD DETECTION IN MEDICAL CLAIM USING DATA MINING TECHNIQUES – A COMPARATIVE STUDY

Anu Victor
Lecturer, Department of Sciences, St. Mary's College, Yousufguda, Hyderabad
Email: anuvictor@stmaryscollege.in

Aparna Vijayan
Assistant Professor, Department of Sciences, St. Mary's College, Yousufguda, Hyderabad
Email: aparnabinod@stmaryscollege.in

Received: Oct. 2019 Accepted: Nov. 2019 Published: Dec. 2019

Abstract: Insurance fraud and its abuse have led to significant additional expense incurred in the health care system. According to a study conducted in 2018, by leading consultancy firm EY on financial crime risks in the Indian insurance sector, it is mentioned that over 56 percent of life insurers surveyed witnessed a 30 percent rise in insurance frauds.

Another survey estimates that the number of false claims in the Industry is approximately 15% of total claims made and the report opines that the Healthcare Industry in India is losing approximately Rs.600-Rs.800 cores on fraudulent claims, annually.

A lot of research has gone into achieving effective fraud detection through sophisticated antifraud approaches incorporating various data mining and machine learning methods. An understanding of the medical claim process reveals the involvement of three entities: the service provider (hospitals, labs etc.); the insurance subscribers (patients and/or their employers) and the insurance carriers (insurance companies who receives the premium). Any intervention to combat heath care fraud in these entities can be classified into 3 categories: prevention, detection and responding to fraud abuse.

For the purposes of this paper, it is proposed to examine the different approaches to **medical claim fraud detection,** falling under 3 classes: supervised - which uses samples of previously known fraudulent and non fraudulent methods; unsupervised - which do not require a prior knowledge of fraudulent behavior and hybrid - which uses a combination of supervised and unsupervised techniques. This paper is a comparative study of few data mining techniques for identifying fraudulent detection.

Keywords: Data Mining, Medical Claim, Fraud Detection

Introduction: A health insurance system provides services that cater to the health care needs of a target population. Unfortunately, fraud in health insurance system is surprisingly common and can be carried out in large scale. A fraud in medical claim is an intention to misrepresent relevant information for the benefits of an individual or a group of people. A good amount of healthcare expenditure is wasted due to fraud and abuse. Hence detection and prevention of fraud is a priority issue for the healthcare systems.

Healthcare insurance fraud involves three parties and they are: Service Providers, Insurance Carriers and Insurance Subscribers. Various data mining techniques can be used to detect and avoid fraudulent transactions which require knowledge of health care system, its fraudulent behaviors and characteristics of health care insurance data. Data mining learning techniques are classified into ***supervised*** and ***unsupervised***. Supervised is the most common learning technique. This model uses predefined class labels. In case of health insurance fraud detection, the class labels can either be a "legitimate" or a "fraudulent" claim. The training dataset are used to build the model. At any time, a new claim can be compared with the already trained model to predict its class. A claim will be classified as a legitimate claim if it follows a pattern that is similar to the legitimate behavior else it will be classified as an illegitimate.

Any new type of fraud cannot be detected in a supervised technique. This requires significant efforts from the experts to derive the labeled training samples that can be used to construct the model. Unsupervised learning has no class labels. Here the focus is on finding those instances which displays unusual behavior. Unsupervised learning techniques can discover both existing and new types of fraud as they are not limited to the fraud patterns which already have pre-defined class labels.

Both techniques have its own set of advantages and disadvantages. One technique alone will not help in predicting or detecting a fraudulent transaction. Hence, there is a need to combine the advantages of both the techniques to form a hybrid approach to effectively detect fraudulent claims in health insurance industry. Some of the common types of frauds in health insurance sector that data mining techniques can detect are [1]:

- Charging excessive prices for a treatment /medicine.
- High number of invoices for a particular insuree in a short period of time.
- Insuree buying medicine without medical examination.
- Claiming medical invoices with dates prior to or after the starting of the insurance period.
- Excessive number of medicine claims in a specific period.

Theoretical Background: Data mining for detecting fraudulent data in healthcare insurance claim is an emerging potential area as a result of discovering new patterns and trends in voluminous data generated by healthcare transactions. Classification and Clustering are the two types of data mining learning methods which characterize data points into groups by one or more features. Following are some of the techniques which are applied in the papers reviewed:

Anomaly Detection: The goal of anomaly detection is to identify cases that are unusual within data that is seemingly homogeneous.. In health care, the anomaly detection technique calculates the probability of each claim to be fraudulent by examining the previous insurance claims. The analysts further investigate the cases that have been flagged as fraudulent by data mining model.

Non-Negative Matrix Factorization: Non-negative Matrix Factorization (NMF), an extraction algorithm which is useful when there are many ambiguous attributes which are then combined to produce a meaningful pattern. Data which are multivariate are decomposed by creating user-defined features. each feature is a linear combination of the original attribute set. The coefficients of these linear combinations are non-negative.

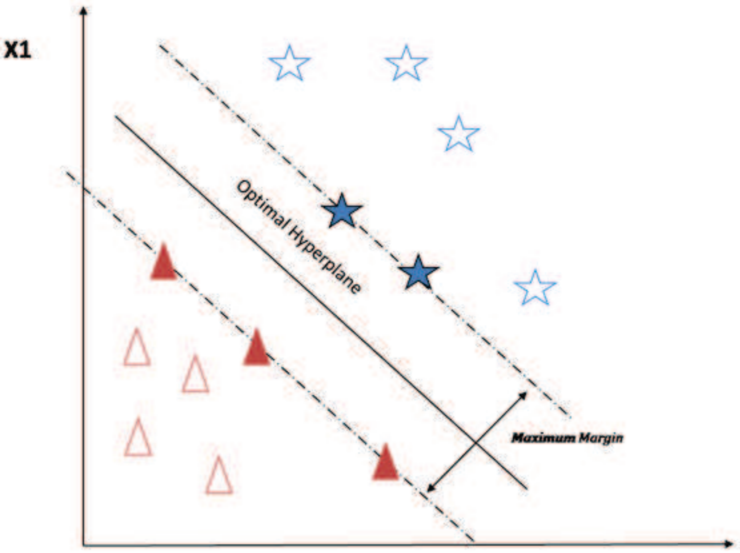
K-Means Algorithm: K-means is a simple and efficient clustering method.Wu et al., (2005) [2] have stated that k-means is one of the best algorithms in data mining. Though it is an efficient algorithm in terms of speed and simplicity, the number of clusters should be known in prior becomes a major drawback for the new incoming objects. Another disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.

Outlier Detection: Outlier detection is a method that helps to identify data patterns that do not match with expected behavior. It is used extensively in a wide variety of applications including insurance or medical care. A baseline of the usual behavior in usage of dosage of medicine for patient is established to detect a fraudulent medical claim and any deviation from the baseline indicates an outlier.

Decision Tree: A decision tree which is a supervised machine learning algorithm is defined by Han and Kamber(2006) [3] as a visual chart that is represented in a hierarchical structure. The root represents the top most node in the tree. An internal node which represents a test on an attribute is generated as a result of applying a splitting algorithm. Among various decision tree algorithms, C5.0 is commonly used because of its advanced mechanisms such pruning level (which allows tuning the severity of tree pruning algorithm), adaptive boosting (which builds a sequence of classifiers and uses a voting strategy to reach the final classification) and misclassification weights (which allows defining different costs for different errors classification) [4]. It was also applied as a divide-and-conquer technique to detect insurance subscribers fraud[5].

Logistic regression: Logistic Regression makes use of logistic function for classification of input datasets. Also called as sigmoid function, it is a S-shaped curve whose input can be any real number. It then maps the input values to a value between 0 and 1. In healthcare, binary logistic regression are used to detect fraud where the classification variable can be true or false for a model or success or failure for a treatment [6]. The logistic regression measures the relation between the dependent variables and the independent variables by estimating probabilities using a logistic function.

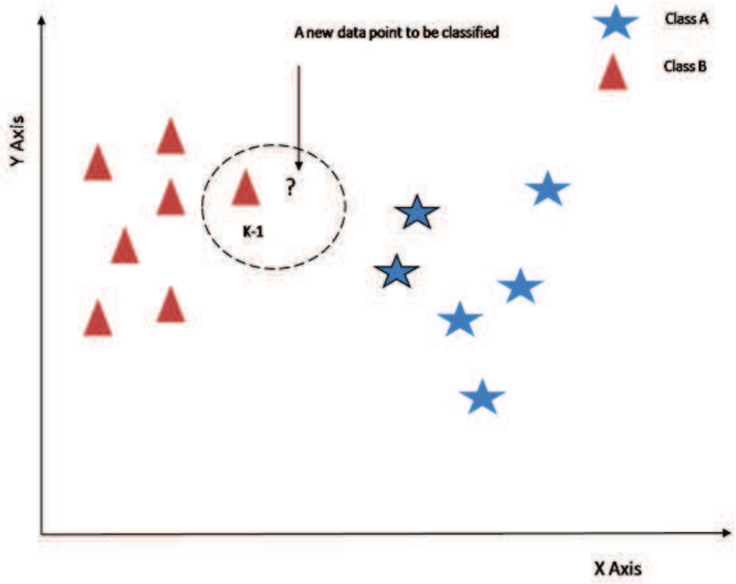
K Nearest Neighbour: The KNN algorithm assumes that similar things exist in close proximity. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness). In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.



MLP Multi Layer Perceptron Neural Networks: An artificial neural network is composed of connected elements, just like neurons, that work together to solve a problem. They are highly useful as they can handle big amount of data having complex structures .Some of the nodes receive scalar data from other nodes and transform the information to a single output signal. The interconnections are weighted and the weights are tunable. This method is useful for detection of fraud in healthcare, due to its enormous variable and quantity of data [7].

Bayesian Co- Clustering Method: A Bayesian Network is a directed acyclic graph that consists of a set of random variables [7]. The Bayesian Network is composed of nodes that are connected by arrows which help the network in representing the dependence between different variables. Each node is associated with a conditional probability table that quantifies the effect of the parent nodes. Bayesian co-clustering method is a powerful data mining tool that can be used to analyze dyadic data connecting two entities where data are represented as a matrix with rows and columns representing each entity respectively[8].

Support Vector Machines (SVM): SVM is a supervised learning model which represent the data as points in space, mapped so that the data points of separate categories are divided by a clear gap that is as wide as possible. The main idea behind the Support Vector Machine is to generate a multidimensional hyper-plane. This hyper-plane can further be used to discriminate between two classes. New data points are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM can be used in detecting medical claim fraud where the system is trained to determine a decision boundary between "legitimate" and "fraudulent" claims. Each claim is then compared with the decision boundary and is placed into the appropriate class. Each insurance claim is then placed into either legitimate or fraudulent class.



Evolving Clustering Method(ECM): It is an unsupervised, clustering method to cluster dynamic data, as the input stream of data is not static. In the clustering process, when a new data point arrives, it is added to a cluster based on its similarities with the existing cluster. When a new cluster is created, the cluster center is identified and its radius is initialized to zero. The radius of the cluster keeps increasing as the new data points are added in. A new data point will not be further added into the cluster when its radius has reached a special value called the threshold value.

Literature Review: [9]The author states that supervised techniques cannot classify new types of disease claims whereas unsupervised techniques cannot detect duplicate claims. Hence a hybrid approach is proposed which consist of Evolving Clustering Method (ECM) for clustering as the data is dynamic and Support Vector Machine (SVM) for classification for detecting health insurance frauds.

[10]This paper develops a model by combining the relevant data sources, then screens the target variable and establishes rules for transformation. Data for the training as well as the verification group is constructed after the selection is made on the time point of the samples. The training group data is used to develop the decision tree model while the verification group evaluates the actual performance of the model.

[11]This paper proposes a Non-Negative Matrix Factorization (NMF) for fraud detection. This technique clusters medical treatment items into several groups according to usage of different patients. Then each group is considered as a kind of medical treatment items for curing similar symptoms. If a medical treatment item shifts between clusters in two consecutive months, then it classifies the patient using this medical treatment item as suspicious patient. In the end, all these fraud suspicious patients are submitted to medical experts for detailed careful detection. The factorization can be used to compute a low rank approximation of a large sparse matrix along with preservation of natural data non negativity. Each vector component is assigned a positive value, if the corresponding medical treatment item is used by the patient, otherwise a zero. This method can identify unknown types of frauds.

[12] Liou et al. (2008) reviewed claims submitted to Taiwan’s National Health Insurance for diabetic outpatient services using supervised methods. The selected nine input variables related to expenses such as average diagnosis fee, average consultation and treatment fees ,average drug cost per day ,average drug cost, average days of drug dispense, average dispensing service fees , average medical

expenditure per day, average amount claimed etc. Three data mining methods such as logistic regressions, neural networks and classification trees were compared to check for the efficiency to detect fraudulent / abusive behavior. They concluded that though three methods yielded accurate results, the classification tree model performs the best with an overall correct detection rate of 99%.

[13]Ekina et al. (2013) made use of Bayesian co-clustering methods to detect patterns within provider-user pairs that may signal potential fraudulent activity. Initially, the claim information is mapped to a visitation matrix which is composed of a collection of indicators that specify whether the users have claims associated with a provider. For example, assume that there are i health-care providers and j health-care service beneficiaries. Let Xij, be a binary random variable which represents the provider i that serves user j. Xij = 1 if provider i serves beneficiary j (0 otherwise). After generating visitation matrix, behavioral patterns are identified to group provider similar services to the similar users using Bayesian co-clustering algorithm .

[14] This paper demonstrates how data mining can be used to identify indicators of healthcare fraud and abuse in general physicians' drug prescription claims submitted to the Social Security Organization (SSO) in Iran. 5 steps involved to detect fraud are: identify the nature of the problem and its objectives, data preparation, identification of indicators and its selection, identify suspect physicians using cluster analysis, and discriminant analysis to assess the validity of the clustering approach. In this process thirteen indicators were developed. Two indicators were related to cost issues, 4 indicators were related to the frequency and patterns of visits, and 7 indicators were related to prescription patterns. The indicator, 'the average cost of a drug prescription claim' was shared in the detection of abuse and fraud. The higher the value of the indicator means that there is a high possibility of abusive or fraudulent behavior.

[4]Liu et. al, have suggested a geo location clustering design where the particulars based on geo location of Medical claim recipient and service providers. In their study, they used cluster analysis to group Medical claims according to claim payment amount and the distance between beneficiary and service provider. This helped in identifying not only the claims with extreme payment amounts and distances but also recognizing some potential outliers that cannot be easily detected when analyzing only claim payment amount or distance. In the data preparation phase, geo location information of countries and state4s was collected which was then mapped to the medical claim data. In the data preprocessing phase, Euclidean distance between beneficiaries and service providers was calculated using the latitude and longitude information mapped to the Medical claim dataset in the data preparation stage.

Claims were then classified based on the principal diagnosis and the Claims having three most common diagnoses were selected. In the analysis phase, the three data sets according to the payment amount and distance was analyzed which was then clustered to compare the results.

[15]In this research paper the sample data collected from Health Maintenance Organization in Lagos, Nigeria contained 2,474 enrollee claims with 8 attributes. Based on two attributes, Total Amount billed and Total Amount approved, the data was classified to identify the output into fraudulent and Non-fraudulent cases. To reduce the effect of scale differences the attributes were normalized for comparison. Six clusters were formed from these two attributes. The claims were then grouped into the clusters as: cluster 0(0%), cluster 1(16%), cluster 2(74%), cluster 3(2%), cluster 4 (2%), and cluster 5(6%). The number of iterations was 49 and number of cluster selected was 6. Within each clusters sum of squared errors was 2.474946333312165 and total number of claims identified as possible anomalies which were outliers was 7.

[16] JiwonSeo et.al (2017) proposed a Page Rank-based algorithm for identifying the frauds and anomalies in Medicare-B dataset. The Medicare-B dataset contains medical insurance claim information of public including 162$ billion and 8880,000 providers and 10 million claims in the dataset. The author selected the personalized Page Rank algorithm to identify anomalies based on the similarities between the providers and their prescriptions. High values of Page Rank imply legitimate treatment by the

provider with a different specialty. This algorithm uses a random walk to select a node and follows the similar nodes with similar prescriptions and converges the seed nodes to have high PageRank values. The algorithm repeats over a time and examines the providers with their PageRank values, if it is smaller, then it is likely to be fraudster or anomaly.

[17] used MLP multilayer to detect health care fraud in Chile. They made use of ten independently trained neural networks for each one of the entities involved in the fraud/abuse problem: medical claims, affiliates, medical professionals and employers. Based on the medical claim received , the different entities are analyzed separately using historical data with cross-references among them. This divide-and-conquer strategy allows to feedback information over time, combining affiliates', doctors' and employers' behavior.

[18]Here each new case is classified based on the classification of its k nearest neighbors.Though the distance from a new case to its nearest neighbor can be calculated in several ways , this study employed genetic algorithm to optimize the feature weights instead of the conventional distance metric, The study compared the classification accuracy obtained when two decision rules (Majority rules and Bayes rule) are used to derive a classification from 1,3,5 nearest neighbor.

[19]The paper proposes a hybrid model for detecting insurance frauds which uses both supervised and unsupervised techniques. The methods chosen are:- K-Means Clustering Method and Support Vector Machine (SVM).Unsupervised techniques K-means is chosen for partitioning datasets. Support vector machines, supervised learning technique, is a classification tool that can be used for distinguishing datasets. Initially, K-means Clustering algorithm is designed followed by building classifiers. Hence combining the advantages of Supervised and unsupervised learning a hybrid model is designed. The numeric dataset is applied for the support vector machines classification. The plotted dataset are compared with one another to get a better visualization of data. These features give a clear picture of all numeric data and their target class.

[20] This paper used a hybrid model that combines unsupervised SOM and supervised MLP neural network was proposed to classify service providers' profile .The training data were originally divided into four classes indicating different likelihoods of fraud. The classification results were not satisfactory when the authors applied only MLP neural networks to the data. Hence, Self organizing maps (SOM)was employed to refine the training data. The SOM indicated that only two classes were well defined by the classification given by human experts. Hence, the two classifications were used to retrain the MLP neural network and this led to better classification results.

[21] This paper proposes Decision Tree (J48, ID3 and NAÏVE BAYES) as a powerful classification algorithm and can be used for the discovery of health care fraud. The data is collected and formed according to the required format and structures and later it is converted to ARFF (Attribute Relation File Format) format to process in WEKA. According to the outcome of the research ID3 has the highest accuracy of 100% and J48 with the lowest accuracy of 96.72%. Hence decision tree ID3 is the best of the three respective algorithms as it is more accurate.

[22]The medical health claim data mostly consists of patient records which may have several different types of features such as patient age, blood group, weight, disease cause, types of diagnosis done etc. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors, wrong diagnosis etc. Outlier detection methods aims at detecting irregular records referred as point outliers. Generally the labeled data belongs to the healthy patients, hence most of the techniques adopt semi-supervised approach. To detect outliers in such data, collective outlier detection techniques can be applied [23]. Various other techniques have also been applied to detect disease outbreaks in a specific area [24]. Hence the outlier detection is a very critical problem in health care domain which requires high degree of accuracy.

Conclusion: Fraud becomes more diverse, as the amount of data grows. Reduction of fraud can result by elimination of fake claims. We reviewed papers on various Data mining techniques to detect fraudulent claims in heath care. During the course of the review it has been observed that neither supervised nor can umsupervised alone give accurate result in predicting the fraudulent medical claims. As a result we also agree that a hybrid model can predict an accurate result. The future scope of this review is to design a hybrid model to identify fraudulent medical claim.

References:

1. Melih Kirlidoga, Cuneyt Asukb , '*A fraud detection approach with data mining in health insurance*', Procedia - Social and Behavioral Sciences 62 (2012) 989 – 994.
2. Hung, Ming-Chuan& Wu, Jungpin & Chang, JH & Yang, Don-Lin. (2005). '*An Efficient k- Means Clustering Algorithm Using Simple Partitioning*', Journal of Information Science and Engineering. 21. 1157-1177.
3. Han, J. and Kamber, M. 2006, '*Data mining: Concept and Techniques*', Second Edition, Morgan Kauffmann Publishers, San Francisco.
4. Qi Liu , Miklos Vasarhelyi, '*Healthcare fraud detection : A survey and a clustering model incorporating Geolocation information.*' (2013), 29th World continuous auditing and reporting symposium (29wcars), november 21-22, 2013, brisbane, Australia.
5. Aranha, Claus & Iba, Hitoshi. (2009)., '*The Memetic Tree-based Genetic Algorithm and its application to Portfolio Optimization. Memetic Computing* ' 1. 139-151. 10.1007/s12293-009-0010-2.
6. F.-M. Liou, Y.-C. Tang and J.-Y. Chen, '*Detecting hospital fraud and claim abuse trough diabetic outpatient services*', in Health Care Manage Sci, 2008.
7. S. Maes, K. Tuyls and B. Vanschoewinkel, '*Credit Card Fraud Detection Using Bayesian and Neural Networks*', in Interactive image-guided neurosurgery, Brussel, Belgium, 2002.
8. Shan, Hanhuai & Banerjee, Arindam. (2008). '*Bayesian Co-clustering*'. Proceedings - IEEE International Conference on Data Mining, ICDM. 530-539. 10.1109/ICDM.2008.91.
9. Rawte, Vipula&Srinivas, Anuradha.(2015), ' *Fraud detection in health insurance using data mining techniques*'.1-5. 10.1109/ICCICT.2015.7045689.
10. Lin, Kuo&Yeh, Ching& Huang, Shih.(2013). '*Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance.Applied Mechanics and Materials*'. 284-287. 1574-1578. 10.4028/www.scientific.net/AMM.284-287.1574.
11. Shunzhi Zhu, Yan Wang, Yun Wu, "Health care Fraud Detection Using Nonnegative Matrix Factorization", The 6th International Conference on Computer Science & Education, 2011.
12. *Liou FM, Tang YC, Chen JY,* 'Detecting hospital fraud and claim abuse through diabetic outpatient services'*Health Care Manag Sci. 2008 Dec; 11(4):353-8.*
13. Ekin, Tahir &Caglar, Toros &Soyer, Refik. (2012). '*Application of Bayesian Methods in Detection of Healthcare Fraud*'. https://pdfs.semanticscholar.org/7f47/e792a8551c1d11635fc38e4ff44e14c5d599.pdf
14. Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). '*Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study*'. *International journal of health policy and management*, *5*(3), 165–172. doi:10.15171/ijhpm.2015.196
15. Fashoto Stephen G., Owolabi Olumide, Sadiku J. and Gbadeyan Jacob A, '*Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm*' Australian Journal of Basic and Applied Sciences, 7(8): 140-144, 2013 ISSN 1991-8178
16. JiwonSeo and OferMendelevitch, '*Identifying frauds and anomalies in the medicare-B dataset*'. IEEE, pp.3664-3667, 2017
17. Ortega PA, Figueroa CJ, Ruz GA (2006) '*A medical claim fraud/ abuse detection system based on data mining: a case study in Chile*'. In Proceedings of International Conference on Data Mining, Las Vegas, Nevada, USA
18. H. He, S. Hawkins, W. Graco, and X. Yao, '*Application of Genetic Algorithm and K-Nearest Neighbour Method in Real World Medical Fraud Detection Problem,*' J. Adv. Comput. Intell. Intell. Inform., Vol.4, No.2, pp. 130-137, 2000.

19. MsJanhavi Naik, Dr J.A Laxminarayana , ‘*Designing Hybrid Model for Fraud Detection in Insurance’*, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 24-30
20. He H, Wang J, Graco W, Hawkins S (1997) Application of neural networks to detection of medical fraud. Expert Syst Appl 13:329–336
21. Rekha Pal and Saurabh, Pal VBS Purvanchal University,*’ Application of Data Mining Techniques in Health Fraud Detection’* , International Journal of Engineering Research and General Science Volume 3, Issue 5, September-October, 2015 ISSN 2091-2730
22. Karanjit Singh and Dr. Shuchita Upadhyaya, ‘Outlier Detection: Applications And Techniques’, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814
23. https://pdfs.semanticscholar.org/d55b/2de12eac42739035e15173a2bf4828b81b.pdf
24. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.450.3026&rep=rep1&type=pdf

***