A Survey of different methods of clustering for anomaly detection

Sarita Tripathy,Prof(Dr.)Laxman Sahoo

Abstract - Anomaly detection is the process of identifying unusual behavior. It is widely used in data mining, for example, to identify fraud, customer behavioral change, and manufacturing flaws, data mining techniques make it possible to search large amounts of data for characteristic rules and patterns .With the ever increasing amount of new attacks in today's world the amount of data will keep increasing and because of the base-rate fallacy the amount the false alarms will also increase. Another problem with detection of attacks is that they usually aren't detected until after the attack has taken information. Most current network intrusion detection systems employ signature-based methods or data mining-based methods which rely on labeled training data. Clustering is now the most widely used technique for intrusion detection.

Index Terms: Anomaly detection, Unsupervised learning,K-means Clustering, Fuzzy C-means clustering, Genetic Algorithm,Non negative matrix factorization,Principal component analysis,Coclustering,ID3 decision tree,Hierarchical based clustering**.**

1 INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not confirm to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, expectations, aberrations, surprises, peculiarities or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection. Anomaly detection finds extensive use in wide variety of application domains, for example, an anomalous traffic pattern in computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. An anomalous MRI image may indicate credit card or identity theft or anomalous readings from a space craft sensor could signify a fault in some component of the space craft. Detecting outliers or anomalies in data has been studied in the statistics community as early as the 19th century, over time, a variety of anomaly detection techniques has been developed in several research communities. Many of these techniques have been specifically developed for certain application domains, while others are more generic.

Clustering basically is the task in which the data points are divided into homogenous classes or clusters. Homogenous means there are similar Items present within the same class which are as much as similar. Thus this process can also be referred to as grouping. Clustering is a popular unsupervised pattern classification technique which partitions the input space into number of regions based on some similarity/dissimilarity metric such that similar elements are placed in the same cluster while dissimilar ones are placed in separate clusters. This survey tries to provide an overview of various clustering methods used for anomaly detection. Reminders of this paper organized as the second section gives an overview of how clustering is useful in anomaly detection. Third section gives a description of different anomaly detection approaches, fourth section describes feature selection and reduction, fifth section gives an overview of different clustering algorithms for anomaly detection, and sixth section is the final conclusion.

2.HOW IS CLUSTERING USEFUL IN ANOMALY DETECTION

Clustering can be used as a technique for training of the normality model, where similar data points are grouped together into clusters using a distance function. Clustering is suitable for anomaly detection, since no knowledge of the attack classes is needed whilst training. Contrast to this other learning approaches e.g classification, where the classification algorithm needs to be presented with both normal and known attack data to be able to separate those classes during detection.

To solve the difficulties, we need for detecting new and unknown type of intrusions. A method that offers promise in this task is anomaly in the data (i.e. data instances in the data that deviate from normal or regular ones).It also allows us to detect new types of intrusions, because these new types will, by assumptions, be deviations from the normal or regular ones).It also allows us to detect new types of intrusions, because these new types will, by assumptions be deviations from normal network usage, just like the other types of intrusions. There are several approaches to anomaly detection. Some use known to be normal and use it as a reference for detecting anomalous data. This approach is an example of supervised anomaly detection, since the classification of data must be known prior to training on it. Methods for supervised anomaly detection do not assume that the data is labeled or somehow otherwise sorted according to classification. One method involves building probabilistic models from the training data and then using them to determine whether a given network data instance is an anomaly or not. Another method is clustering similar data instances together into clusters and use distance metrics on clusters to determine what is an anomaly. Clustering can be performed on unlabeled data, requiring only feature vectors without labels to be presented. There are several primary assumptions having the same classification (type of attack or normal) should be close to each other in feature space under some reasonable metric, while instances with different classification will be far apart.

Clustering is a well known and studied problem. It has been studied in many fields including in statistics[1], machine learning[2],database[3],and visualization. Basic methods for clustering include the linkage based[4] and k-means[5] technique.k-means makes several passes through the training data and each pass shifts cluster centers to the mean of data points assigned to that cluster. It then reassigned data points to the nearest prototype, and continues iterating in this manner until no significant changes in the cluster center position occur. The k-means method generally produces a more accurate clustering than linkage based method, but it has a greater time complexity and this becomes an extremely important factor in network intrusion detection due to very large datasets exists, they still do not perform sufficiently fast for datasets with high dimensionality. Most current anomaly detection system estimate the probability based on training the self similar behavior in the internet, in other words events do not occur at a certain rate during any time scale. Therefore, it is wrong to predict the average rate of an event over time window. Clustering and related techniques can find outliers in a dataset.Knorr et al [7] defined that an outlier is an object far away from most objects in the dataset. Instead of using a global density parameter,LOF[8] used a local variable and found local outliers with respect to the local regions. However, these methods parameter is hard to choose and the approaches are not good for high dimensional data such as network data.ramaswamy et al.[9] described an outlier by distance of the kth-nearest neighbor and aggarwal et al[10] calculated the sparsity comparing with the observed and expected number of data in spatial grid cells. A large number of clustering algorithms exist [11,12],but it is difficult to find a single clustering algorithm to get well detection effect. Mean while, attacks are being more diversified, distributive and comprehensive, and detection environmental is also variable. So it is wise to apply several different clustering detection algorithms to given data and then determine the best algorithm for data. Based on this, clustering ensemble combines different algorithms or the same algorithm with different parameters to get better result compared with the single algorithms.

2.1 Clustering: An Unsupervised method for anomaly detection

Clustering analysis is a kind of unsupervised study method[3],some unknown modes are divided into some classes, if some modes have equal distance of characteristic vector within the scope of certain error margin, then they are regarded as the same class. This method can be carried on unlabelled data; it divides the similar data to same class, divides the dissimilar data to different class. Unlike other data mining method, unsupervised anomaly detection method doesn't depend on data class determined in advance, or training sample set with class mark. It doesn't need to input complete normal data while carrying on training, just needs to provide network original data or system audit data with preprocessed simply. This method can find abnormal data having standard, or targeted automatically from data set, in which we don't know what is normal, or what is abnormal. Within the term of unsupervised anomaly detection method, the meaning of "supervised" is having standard, or target variable. Unsupervised method doesn't know the characteristic of each class in advance; summarize it after applying unsupervised method. Here, unsupervised algorithm has two rules for data set. One is that the record number of normal activities must be bigger than record number of intrusion event the other is that there is essential difference between normal record and abnormal record.

3. Anomaly detection approaches

The approaches used to address the anomaly detection problem depend on the nature of the data that is available for analysis. Many approaches exist for anomaly detection. We can roughly classify network anomaly detection into three groups: classification, spectral analysis [13], and clustering. Clustering methods classify data based on similarity based on distance functions such as Euclidean. A good set of clusters should have intra-similarity and inter-similarity .As an algorithm framework for data analysis and interpretation, clustering has been widely used in understanding data, revealing fundamental phenomena, and visualizing major tendencies. Clustering can be categorized into two types: hierarchical clustering uses previously established clusters to find successive clusters and partitioning clustering determines all clusters at once based on iterative procedures. Subspace clustering, correlation clustering and biclustering are new emerging algorithms and have been adapted to practical applications (Density based spatial clustering of application with noise),and probabilistic model-based techniques, such as DBSCAN(density based spatial clustering of application with noise),and probabilistic model based techniques, such as ,Auto class and K-means clustering have also become popular.

4. Feature selection and reduction

As pointed out by jain et al[14,15] and bishop[16],feature selection chooses distinguishing features from set of candidates, while feature extraction utilizes some transformation to generate useful and novel features from the original ones. Both are very crucial to the effectiveness of the clustering applications. Elegant selection of features can greatly decrease the workload and simplify the subsequent design process. Generally, idea features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret. More information on feature selection can be found in{16],and [17].Often many of the dimensions in a data set-the measured features-are not useful in producing a model. Features may be irrelevant or redundant. Regression and classification algorithms may require large amounts of storage and computation time to process raw data, and even if algorithms are successful the resulting models may contain an incomprehensible number of terms. Because of these challenges, multivariate statistical models often begin with some type of dimension reduction. Dimension reduction often leads to simpler models and fewer measured variables, with consequent benefits when measurements are expensive and visualization is important. Feature selection is preferable to feature transformation when the original units and meaning of features are important and the modeling goal is to identify an influential subset.

When categorical features are present, and numerical transformations are inappropriate, feature selection becomes the primary means of dimension reduction.Nonnegtive matrix factorization and principal component analysis (PCA) are widely used techniques for feature transformation.

4.1 Nonnegative matrix Factorization (NMF)

In the real world, the data we used are not often exact because of the related devices 'limited bandwidth, noise, and other degradations. The actual information contained in the original data expressed by many dimensions (features) might be overlapping and interrelating, because it is less precisely defined. Feature selection to select independent and not interrelated variables, feature reduction to get their low-rank approximation and reduce the computation complexity for huge databases, and feature transformation to combine different variables through linear or nonlinear conversion and form remarkable features are necessary in most cases. Given a nonnegative m x k and k x n matrices W and H,respectively,that minimize the norm of the difference X-WH.W and H are thus approximate nonnegative factors of X.The K columns of W represent transformations of the variables in X the k rows of H represent the coefficients of the linear combinations of the original variables in X that produce the transformed variables in W.Since k is < the rank of X,the product WH provide a compressed approximation of the data in X.The possible values for K are often suggested by the modeling context.

4.2 Principal component analysis (PCA)

PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into set of values of linearly uncorrelated variables called principal components. Each principal component is a linear combination of the original variables. This transformation is defined in such a way that the first principal component is a single axis in space .When you project each observation on that axis, the resulting values from a new variable .The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable.PCA is sensitive to the relative scaling of the original variables. Define a data matrix, XT, with zero empirical mean, where each of the n rows represents a different repetition of the experiment, and each of the m columns gives a particular feature. The singularity value decomposition of X is X=W∑VT, where the mxm matrix of eigen vectors of XTX.The PCA transformation is then given by: YT=XTW.if we want a reduced dimensionality representation, we can project X down into the reduced space defined by only the first L singular vectors, WL: Y= (WL) T X=∑LVT, ∑L is rectangular identity matrix. The matrix W of singular vectors of X is equivalently to the matrix W of eigenvectors of the matrix of observed covariance C=XXT.

5. VARIOUS CLUSTERING APPROACHES FOR ANOMALY DETECTION

5.1 K-means algorithm for anomaly detection

The k-means algorithm[19] groups N data points into K-disjoint clusters where K is a predefined parameter such that K<N,The steps in the K-means clustering-based anomaly detection method are as follows:

1. Select K-random instances from the training data subset as centroids of clusters C1,C2…..Ck
2. For each training instances X:
   a. Compute the Euclidean distance:
      D(Ci,X), i=1,…..k.
   Find cluster Cq that is closest to X.
   b. Assign X to Cq. Update the centroid of Cq(the centroid of a cluster is the arithmetic mean of the instance in the cluster)
3. Repeat step2 until the centroids of cluster C1, C2, .Ck stabilize in terms of mean squared error criterion. Finally, the algorithm aims at minimizing an objective function.

This clustering algorithm is data driven method which has relatively few assumptions on the distribution of the underlying data. Besides its greedy search strategy guarantees at least a local minimum of the objective function, thereby accelerating the convergence of cluster on huge ARP (Address Resolution Protocol) traffic. This leads to higher performance for the proposed algorithm comparing to other Anomaly detection algorithm.

5.2 Integration of Fuzzy C-means and Genetic Algorithm for anomaly detection

Genetic Algorithm method with data reduction is applied to identify subset of features for network security and fuzzy C-means[18] is used for clustering group of data using fuzzy c-means clustering, it is used to classify between normal class and certain intrusion class. Here genetic Algorithm will be used for feature selection and reduction which aims at selecting a small or prespecified number of features leading to the best possible performance of the entire classifier. Therefore, the main goal of feature subset selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy.

5.3 Adaptive hierarchical based clustering for anomaly detection

Traditional anomaly detection methods lack adaptivity in complex and heterogeneous network. Especially while facing high noise environments or situation of updating profiles not in time, intrusion detection systems will have high false alarm rate. The improved hierarchical clustering tree is developed in[] which supports updating profiles at any time. Here the clustering algorithm is extended and branch and bound mechanism is applied for filtering noise. With the help of two advantages: filtering noise and updating profiles at any time, our algorithm is effective enough to meet adaptive requirements.

5.4 K-means+ID3 for anomaly detection

The proposed method has two phases: 1) training and 2) testing, during training phase k-means based training method is first applied to partition the training space into k-disjoint clusters C1, C2, C3….Ck then ID3 decision tree is trained with instances in each k-means cluster. The k-means clustering method ensures that each training instance is associated with only one cluster. The testing phase of the algorithm includes two phases:1) the candidate section phase and 2)the candidate combination phase. In the first phase the anomaly scores for the k-means and decisions for ID3 decision tree are extracted. The combined application of the two algorithms overcomes some limitations of each algorithm when applied individually. For example, selection of a right value for parameter K in the K-means clustering algorithm can affect on the overall accuracy of the algorithm. Considerably little values of K, compared to internet number of natural sub groupings within the training data will load to overlapping subgroups within cluster. This problem is compensated by ID3 decision tree constructed in each cluster.

5.5 Coclustering

Both SMR co-clustering and information theoretic co-clustering proved to be powerful tools for detecting anomalous connections out of a large data set without training the algorithms.SMR co-clustering could not compute clusters over the full data set but is effective in finding parameters that are strong indicators of abnormality in the data set and produces clusters that are highly pure. Information theoretic co-clustering[20] on the other hand does not isolate certain parameters consistently and does not always produces pure clusters, but can run on the full data set and be used to plot the full set of connections, against the parameters calculated by SMR co-clustering, colored similarly to that by the labels. Therefore, by combining both methods.co-clustering could be strong technique for system administrators to review network connections, searching for reviewing anomalies.

CONCLUSION

In this paper a brief overview of anomaly detection using clustering is given, some of the clustering techniques which are mentioned in this paper are k-means, Fuzzy C-means, hierarchical based clustering and co clustering. From the study it is concluded that clustering is an immensely efficient approach for anomaly detection and overcomes many limitations of traditional methods. Further work in this field can be done to discover more improved methods.

REFERENCES:

[1] A.ghosh and A.shwatzbadr.A study in using neural network for anomaly and misuse detection.

[2] R.rojas neural networks-A systematic introduction.Springer,Berlin.1996.

[3] Alexander hinneburg and Daniel A.Kim.Clustering methods for large databases.from the past to the future.In Alex delis,Christosn falousos,and shahram ghandeharizadeh editors.sIGMOD 1999,proceeding ACM SIGMOD International conference on management of data,june1-3,1999,Philadelphia,Pensylvania,USA,ACM press 1999.

[4] H-H BOCK Automatic classification vandenhoeck and Ruperech,1974

[5] E.ESkin.Anomaly Detection over noisy data using learned probability distribution In proceedings of International Conference on Machine Learning 2000.

[6] A Ghosh,A Schwartzbard,& M Schatz,learning program behavior profiles for intrusion behavior profiles for intrusion detection proc.1st USE NIX Workshop on Intrusion Detection and N/w monitoring 1999.

[7] E,KNorr,&T.Ng,Algorithms for mining distance –based outliers in large datasets,proc.VLOB,1998.

[8] M.Breunig,H.Kriegel,&R.Ng,Lof:Identifying density-based local outlier,Proc.SIGMOD,2000.

[9] S.Ramaswamy,R.Rastogi,&K.Shim,Efficient algorithms for mining outliers from large data sets,Proc.SIGMOD,2000.

[10] C.Aggarwal& p.Yu,Outlier detection for high dimensional data,proc.SIGMOD,2001.

[11] A.K.Jain,M.N.Murty,and P.J.Flynn,”Data Clustering:A Review”. ACM Computing Surveys,Vol.31,no.3,pp.264-323,sept 1999.

[12]

R.O.Duda,P.E.Hart,and D.G.stork,”pattern Classification”,Seconded Wiley,2001.

[13] M.Thottan,and C.Ji,”Anomaly detection IP networks”,IEEE trans.On Signal processing,51(8),2003

[14]A.jain M.Murty,and p.Plynn,”Data Clustering:A review”,ACM comput.Surv.Vol31.no.3,pp.264-323,1999.

[15]D.Jiang,C.tang and A.Zhang,”Cluster analysis for gene expression data:A survey”,IEEE trans.Knowl.Data Eng,vol.16,n0.11,pp.1370-1386,no.2004.

[16] c,Bishop,Neural networks for pattern recognition.New York.Oxford univ Press 1995.

[17]Handbook of pattern recognition and compiler vision,C.Chen,L.Pau,and P.Wang,Eds.World scientific,Singapore,1993,pp61-124 J J.sk lanky and W.siedlechi,Large scale feature selection.

[18] Witch Chimplee,Abdul Hannan,Abdullah,Md sap,Siripor chimplee and Surat srinoy,”Integrating Genetic Algorithm and Fuzzy C-means for anomaly detection’,IEEE indicon 2005 conference, Chennai India,11-13 Dec,2005 pp-575-579.

[19]MengJiang Liang Shang Haiken BianLing Department of computer science and technology,college north china Electric power University Hebei,Baoding,”Application of intrusion detection based on K-means clustering Algorithm” 2009-International forum on Information Technology and application.

[20]Evangelos E.papallexakis,Alexbeute,Petessteenkiste Department of Electrical and computer

Engineering School of Computer Science Carnegie Mellon University Pittsburg,PA,USA,"Network Anomaly Detection using Co-clustering".2012 IEEE/ACM International conference on Advance in Social networks Analysis and mining.PP-403-410