SURVEY ON OUTLIER DETECTION TECHNIQUES USING CATEGORICAL DATA

K.T.Divya[1], N.Senthil Kumaran[2]

[1]Research Scholar, Department of Computer Science, Vellalar college for Women, Erode, Tamilnadu, India
[2]Assistant Professor, Dept. of Computer Applications, Vellalar College for Women, Erode, Tamilnadu, India

---***---

Abstract-*Anomaly detection is used for identification of items, events or observations which do not conform to an expected pattern or other items in dataset .anomalies are also referred to as outliers, novelties, noise, and deviation and expectation three broad category of anomaly detection. Supervised, unsupervised and Semi supervised anomaly detection .Unsupervised anomaly detections detect anomalies in an unlabeled test data set under the that the majority of the instances in the data set considered normal by looking for instances that seem to fit least to the remainder of the data set. In this paper we bring out the key approaches of anomaly detection*

Keywords—Anomaly detection, Supervised Semisupervised,Unsupervised Anomaly detection, Categorical data

Introduction

Data mining extracts secret and useful information from the data. Previously unknown, useful and high quality knowledge can be discovered by data mining. Outlier detection is an important process in data mining .Outlier detection contains many important applications was deserves more attention from data mining community .outlier detection is an important branch in data preprocessing and data mining as this stage is required in elaboration and of data coming from many application fields such as transportation ecology, industrial process, climatology and public safety. Outliers arise because of instrumental error, human error, natural deviations in populations, fraudulent behavior, and changes in behavior of systems or faults in systems. In the presence of outliers many data mining and machine learning algorithms and also the techniques for statistical analysis might not work well. Accurate removal of outliers may highly enhance the working process of statistical and data mining algorithms and techniques. As considering learning task, anomaly detection may be semi supervised or unsupervised. In semi-supervised anomaly detection task we assume all the training instances come from normal class and also our goal is to be distinguish future instances that come from a different anomalies.In unsupervised anomaly detection task we are provide one sample that is to be a mixture of normal and anomalous instances and the goal is to differentiate them. The proportion of anomalies is small, but it exact values is often unknown and varies from task to task

I. OUTLIERS OR ANOMALIES

Outliers can defined as any data value that seems to be out of place with respect to the rest of data. Numerous definition have been proposed for outlier in data mining such as outlier is an anomaly observation or outlier is one that appears from other members of sample in which it occurs. Another definition for outlier is an observation that deviates very much from other observations. Anomalies are patterns in data that do not conform to a correctly defined notion of normal behavior Fig:1 illustrate anomalies in a simple 2-dimensional data set .The data had two regions, N1 and N2, Since most observations lie in these two regions .points that are sufficiently far away from the regions, example points o1 and o2 and points in the region o3 are

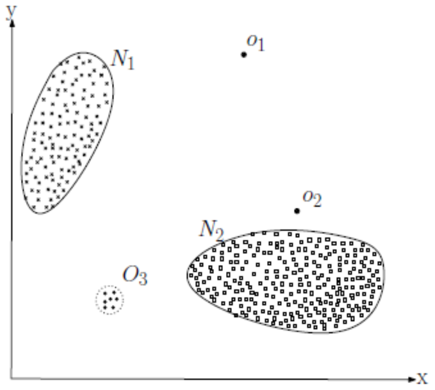


Fig 1: simple example of anomalies in a 2-dimensional data set

Anomaly detection technique is related to, but distinct from noise removal and noise accommodation both are deals with removal of noise

Challenges:

*Defining a normal region which encompasses each and every possible normal behavior is very difficult and also the boundary between anomalous and normal behavior is often not precise.

* Anomalous observation which lies near to the boundary can actually be normal, and vice versa

*When anomalies are the result of vicious actions the vicious adversaries often adapt themselves to make the anomalous observations appear like normal, making the task of defining normal behavior more difficult

*Many domains normal behavior keeps progressing and a current idea of normal behavior might not be sufficiently representative in the future

*Availability of labeled data for validation /training of models used by anomaly detection technique s is usually a major issue

*Often the data contains outlier which trends to be very similar to the actual anomalies and hence a difficult to distinguish and remove

*Because of the above challenges the outlier detection problem, in most general form, is difficult to solve. Most of the existing anomaly detection techniques solve particular formulation of the problem. The formulation is induced by various factor such as , availability of labeled data, nature of data and types of anomalies to be detected etc.Often these factors are calculated by the application domain in which anomalies be detected

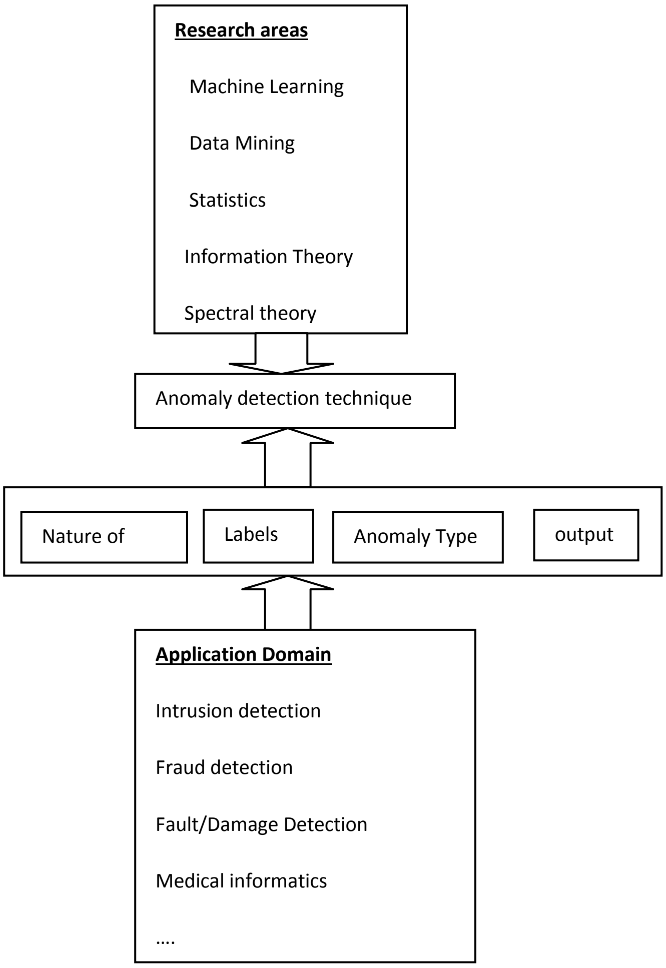


Fig 2:Important components associated with anomaly detection techniques

II. OUTLIER DETECTION APPROACHES

Outlier detection is used to find for objects that do not satisfies rules and expectations valid for the important part of the data. The detection of outlier object may be an proof that there are new techniques in data. outliers are considered errors or noise, they may have important information. The meaning of an outlier is often depend on the applied detection methods and unknown assumptions regarding data structures used. Based on these approaches used in outlier detection the methodologies can be classified as:

A. Statistical outlier detection
B. Depth based outlier
C. Deviation based outlier
D. Distance based outlier
E. Density based outlier

III. VARIOUS OUTLIER DETECTION TECHNIQUES

Outlier detection varies in accordance with different entities in different domains. Formulation of outlier detection is based upon various factor such as input data type and distribution availability of data and resource constraints introduced by application domain. There are outlier detection techniques widely used over streaming data they are as follows

A. *Statistical Outlier detection*

This uses certain kind of statistical distribution and computes the parameters by assuming all data points have been generated by a statistical distribution. Here outliers are points that have a low probability to be generated by overall distributions Statistical outlier detection technique is also known as the parametric approach. This technique is formulated using the distribution of data point available for processing .Detection model is designed to fit the data with reference to distribution of data .Gaussian mixture model was proposed by yaminishiet [1].where each point is given a formulated score and data points which have a high score declared as outlier. Detecting outlier based on general pattern within data points was designed by [2] where it merge both Gaussian mixture model and supervised method

B. *Depth based outlier detection*

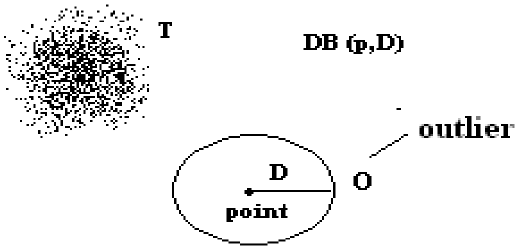
This is one of the variant of statistical outlier detection. Depth based outlier detection search novelty at the border of the data space but which are independent of statistical data Distributions. This techniques are generally suited quantitative real valued data set or quantitative ordinal distributions. Here each data object of dataset represented by an n-D space having a assigned depth such data points are organized into convex hull layers according to assigned depth and anomaly is formulated on the basis of shallow depth values. Anomaly are object on outer layers. These models are not fit for high dimensional data set

C.Deviation based outlier detection

In deviation based outlier detection a set of data points is given. Here outliers are detect as point that do not fit to the general characteristics of the set. so the difference of the set is minimized when removing the outliers .Data elements distributed as like a sparse matrix in data set. This will creates confusion over data analysis .some points are get deviated from normal points are declared as outliers. Sequential problem approach was proposed[4] in where outliers are detected by using normal features of data points and deviated features of data. To deal with time series method oriented information ,jagadish et al proposed a histogram based approach[4]. while considering this method not suitable for streaming data in distributed environment and over multivariate data is left as open.

D.Distance based outlier

This outlier detection technique judge a point based on the distances to its neighbors. Basic model of distance based outlier is given

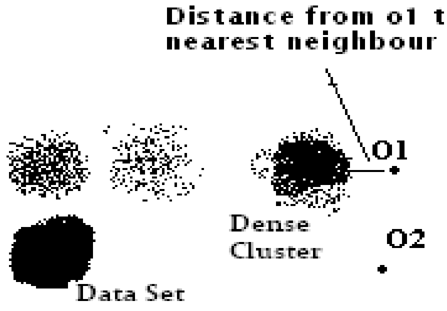


Explicit distance based approaches are depend on the well known nearest neighbor principle.Ng and knor propose[5] a well known distance metric to detect outliers. They define outlier as the object which was greater in distance to its neighbors. The nested loop(NL) algorithm, calculates the distance between each and every pair of objects and then set as outliers those that are far away from most objects. The NL algorithm has difficulties with respect to the number of objects, making it not suitable for mining very large databases such are found in government audit data, network data, and clinical trial data. This outlier method was presented in knorr and Ng as an object O in dataset T is a DB (p,D) outlier if at least fraction p of the object in T lie at a distance greater than D from O.The Parameter p used here is the small fraction of objects that must present outside an outliers D-neighborhood.

This method[6] is further elaborated where they give a prior consideration on distance of a point from its k the nearest neighbor. where the top k point are considered as an outliers. This approach was proposed by Angiulli and pizzuti[7] on the basis of outlier factor. Every data point is assigned formulated outlier factor computed as sum of distance from its k nearest neighbors. For detecting outliers linear time is made use in[8] where data set get randomized for efficient search space .Recently we find out that a non parametric unsupervised based methods used for outlier detection which s proposed by a branch et al[9].To address the uncertainty temporal relation and transiency present within which data distance based outlier detection for data stream method proposed with the help of continuously adaptive data distribution function[10]

F. Density based outlier detection

This method compares the density around point with its local neighbors densities. The relative density of point compared to its neighbors is calculated as an outlier score. This outlier detection method uses density distribution of data points into the data set. The taught of density based local outlier using comparison with density of local neighborhood was introduced by Brewing et al[11].Here an outlier are measured by using a local outlier factor(LOF),which is ratio of local density of point and the local density of its nearest neighbor. Data Point whose LOF value is high is declared as outlier



Data set and dense cluster with outliers

IV. OTHER RELATED WORKS IN ANOMALY DETECTION

Anomaly detection have been the topic of number of surveys and review articles, as well as books. Hodge and Austin[2004] Given an extensive survey of anomaly detection techniques created in machine learning and statistical domains. A complete review of outlier detection techniques for numeric as well as symbolic data is presented by Agyemang et al.[2006].A full review of novelty detection techniques using neural networks and statistical approaches has been presented in Marko and Singh[2003a] .Patcha and park[2007] and sunder[2001].Present a survey of anomaly detection techniques used specially for cyber intrusion detection. Outlier or anomaly detection has always attracted a lot of research interest since its first definition in the late 1960s[9].with the advent of data mining and the advances in machine learning that occurred in the 1990s, research on anomaly detection gained new impetus and gave rise to many novel approach algorithm

V. CATEGORICAL DATA

Categorical data also known as nominal or qualitative multistate data has become increasingly common in modern real world applications. Table 1 shows a sample of categorical data set. These dataset are often rich in information and are frequently detected in domains where large scale data sets are common,e.g.,in network intrusion detection .however ,unlike categorical data, continuous data attribute values cannot be naturally mapped on to a scale, making most continuous data analysis techniques in applicable in this setting.

Cap-shape	**cap-surface ……**	**habitat**	**class**
Convex	smooth	urban	poisonous
Convex	smooth	grasses	edible
Bell	smooth	meadows	edible
Convex	scaly	urban	poisonous
Convex	smooth	grasses	edible

Table 1: mushroom data set

This gave an idea of the range and overall distribution of each attribute .However in categorical data we can only notice at the mode or an unordered histogram. With normal data we may also be able to look at percentiles but for the most part the situation is similar to categorical data

VI. RELATED WORKS IN CATEGORICAL DATA

Data with categorical attributes have been studied for a very long time, dating back at least a century when Karl Pearson[12,13] introduced the test for independence between categorical attributes. The traditional exploratory techniques used were contingency tables, the chi square statistics, pie charts and unordered histogram. Friendly proposed several sophisticated statistical techniques which are Sieve diagrams and Mosaic Displays to view k-way contingency tables and Multiple Correspondence analysis(MCA) to deal with multivariate categorical data sets, though most techniques are limited to attributes that take few possible values. Fernandez discusses several exploratory techniques for categorical data from data mining perspective. There was number of studies directed at categorical data in visualization community. In particular, one in visualization has been to order the categories using the information present in data. one such technique is called Distance Quantification classing(DQC) was developed by Rosario et al to order the categories present in a class variable in categorical data set with respect to the Predictor variable

VII. RESULTS AND DISCUSSION

Outlier detection is an important process in data mining. Outlier detection as a branch of data mining have many important applications and deserves more attention from data mining community. Development of a model that accurately represents the data is required for accurate outlier detection. Past decades, a large number of techniques have been created for building such models for outlier and anomaly detection. To present accurate and effectiveness for outlier detection that should able to handling following weakness with respective outlier detection techniques

Approach	Outlier detection
Statistical outlier detection	78%
Depth based outlier detection	85%
Deviation based outlier detection	80%
Distance based outlier detection	84%

Table 2: Comparison of outlier detection Methods

VIII. CONCLUSION

A large number of techniques have been developed in outlier detection area, but most of them have been some inherent limitations. Outlier detection over streaming data was an important research problem in data mining community. Finding out outlier is important because it contains useful information which may lead for further researching domain. This Paper provides a review of outlier detection methods over streaming data with data mining perspective. Based on review we can conclude that the most of the techniques used are focuses over algorithms. These require a special background and notion of finding anomaly also varies from domain to domain. It is observed that efficiency of outlier detection method is highly based on data distribution and type of data. Some techniques mentioned in this paper require a prior knowledge about data. For instance that statistical technique uses a data distribution and model. Also the assumption made about data is correct. The individual methods are not efficient over streaming data. In such case if prior information about data is not known then better to make use of combine approach for outlier detection

REFERENCES:

[1]K.Yamanishi et al,*2004.On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms*. In Proceedings of Data Min.knowledge Discovery.Vol.8,No 3,pp 275-300

[2] K.Yaminishi and J.Takeuchi,2001.*Discovering outlier filtering rules from unlabeled data combining a supervised learner with an unsupervised learne*r. In proceedings of KDD'01, pp389-394

[3]R.Nuts and P.Rouseeuw, 1996.Computing *depth counters of bivariate point clouds*. Computational Statistics and Data Analysis,Vol 23,No 2,PP 153-168

[4]H.V.jagadish et al,1999.Mining deviants in a *Time series Database.In Proceedings of 25 international conference on very large data bases*.Edinburgh,Scotland,PP 102-113

[5] Knorr E.M,Ng,R.T.,"*Finding Intentional Knowledge of Distance-Based Outliers*", Proceedings of the 25th international Conference on Very Large Data Bases,Edinburgh,Scotland,pp.211-222,September 1999

[6]Ramaswamy S.,Rastogi R.,Kyuseok s:Efficient Algorithms for Mning Outliers from Large Data Sets Proc.ACM SIDMOD Int.Conf.on Management of Data.,2000

[7] F. Angiulli and C. Pizzuti, 2002. *Fast outlier detection in high dimensional spaces*. In Proceedings of PKDD'02, 2002.

[8] Bay S. D. and Schwabacher M., 2003. *Mining distance-based outliers in near linear time with randomization and a simple pruning rule*. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 29-38.

[9] J. W. Branch et al, 2006*, In-network outlier detection in wirelesssensor networks,* In 26th IEEE International Conference on Distributed Computing Systems (ICDCS'06), pp 49.

[10] Sadik, S. and Gruenwald, L. 2010. DBOD-DS: *Distance Based Outlier Detection for Data Stream.* DEXA' 10.

[11] C. C. Aggarwal and P. S. Yu., 2001. *Outlier detection for high dimensional data.* In Proc. 2001 ACM-SIGMOD Int.Conf.Management of Data (SIGMOD'01), pp37-46.

[12] K. Pearson. On the Theory of Contingency and Its Relation to Association and Normal Correlation.Dulau and Co., 1904.

[13] K. Pearson. On the general theory of multiple contingency with special reference to partial contingency. Biometrika, 11(3):145-158, 1916