IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Fraud Detection Using Outlier Analysis: A Survey

Vidya Mohanty[*1], P.AnnanNaidu[2]

[*1, 2] Centurion University, India

vidya.mohanty@gmail.com

Abstract

Outlier detection is a primary step in many data-mining applications. There are several methods for outlier detection, like univariate vs. multivariate techniques and parametric vs. nonparametric procedures. In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for data mining isoften based on distance measures, clustering and spatial methods. An outlier is an observation (or measurement) that is different with respect to the other values contained in a given dataset. Outliers can be due to several causes. The measurement can be incorrectly observed, recorded or entered into the process computer, the observed datum can come from a different population with respect to the normal situation and thus is correctly measured but represents a rare event. An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism.

Keywords: Outliers, Distance measures, Statistical Process Control, Spatial data.

Introduction

In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis.

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. Barnet and Lewis indicate that *an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs*, similarly, Johnson defines an outlier *as an observation in a data set which appears to be inconsistent with the remainder of that set of data*. Other case-specific definitions are given below.

Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks.

Traditional Approaches

The salient traditional approaches to outlier detection can be classified as either distribution based, depth-based, clustering, distance-based or density-based.

Distribution-based method

These methods are typically found in statistics textbooks. They deploy some standard distribution model (Normal, Poisson, etc.) and flag as outliers those data which deviate from the model. However, most distribution models typically apply directly to the future space and are univariate i.e. having very few degrees of freedom. Thus, they are unsuitable even for moderately high-dimensional data sets. Furthermore, for arbitrary data sets without any prior knowledge of the distribution of points, expensive tests are required to determine which model best fits the data, if any! Fitting the data with standard distributions is costly and may not produce satisfactory results. The most popular distribution is the Gaussian function.

A method was proposed by Grubbs which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the attribute, where the mean and the standard deviation are calculated from all attribute values including the query value. The Z value for the query is compared with a 1% or 5% significance level. The technique requires no pre-defined parameters as all parameters are directly derived from the data. However, the success of this approach heavily depends on the number of exemplars in the data set. The higher the number of records, the more

Statistically representative the sample is likely to be. A Gaussian mixture model and computed outlierness was proposed based on how much a data point deviates from the model.

The GMM is represented by equation (1):

$$\mathrm{P(t|x)}=\sum_{j=1}^{M}\alpha_j\,(\mathrm{x})\mathrm{j_j}\,(\mathrm{t|x}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(i)}$$

Where M is the number of kernels (ϕ), $\alpha_j(x)$ the mixing coefficients, **x** the input vector and **t** the target vector.

A Gaussian probability density function is defined by equation (2):

$$\phi_j(t\,|\,x)=\frac{1}{2\pi^{\frac{d}{2}}\sigma_j^{\,d}(x)}e^{\{\frac{\left\|t-\mu_j(x)\right\|^2}{2\sigma^2{}_j(x)}\}} \quad \ldots\ldots\ldots \text{(ii)}$$

Where d is the dimensionality of input space, σ is the smoothing parameter, $\mu_j(x)$ represents the centre of jth kernel and $\sigma^2{}_j(x)$ is the variance.

Clustering

Clustering is a basic method to detect potential outliers. From the viewpoint of a clustering algorithm, potential outliers are data which are not located in any cluster. Furthermore, if a cluster significantly differs from other clusters, the objects in this cluster might be outliers. A clustering algorithm should satisfy three important requirements:

- Discovery of clusters with arbitrary shape;
- Good efficiency on large databases
- Some heuristics to determine the input parameters.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is based on minimization of the following objective function:

$$J_m=\sum_{i-1}^{N}\sum_{j-1}^{C}u^m_{ij}\left\|x_i-C_j\right\|^2 \ldots\ldots\ldots\ldots\ldots\ldots \text{(iii)}$$

where C is the total number of clusters, N is the total number of data, m is any real number greater than 1, uij is the degree of membership of xi to the j-th cluster, xi is the i-th of the d dimensional measured data, cj is the d-dimensional center of the cluster, and || * || is any norm expressing the similarity between any measured data and the center. However, since the main objective of a clustering algorithm is to find clusters, they are developed to optimize the outlier detection. The exceptions (called "noise" in the context of clustering) are typically just tolerated or ignored when producing the clustering result. Even if the outliers are not ignored, the notions of outliers are essentially binary and there are no qualifications as to how outlying an object is.

Outlier Detection Using Artificial Intelligence Techniques

When dealing with industrial automation, where data coming from the production field are collected with different and heterogeneous means, the occurrence of outliers is more the rule than the exception. Standard outlier detection m2ethods fail to detect outliers in industrial data because of the high dimensionality of the data. In these cases, the use of artificial intelligence techniques has received increasing attention in the scientific and industrial community, as the application of these techniques shows the advantage of requiring poor or no *a priori* theoretical assumption on the considered data. Moreover their implementation is relatively simple and with no apparent limitation on the dimensionality of the data.

Neural networks

In 1943, McCulloch and Pitts introduced the idea of an artificial neuron to process data. In the 50ies this work was advanced by arranging neurons in layers. Although learning rules to cope with multiple layers of perceptrons were not developed until later, this work formed the basis of the Multi-Layer Perceptron (MLP) that is used today.

Another kind of neural network that is frequently used is the Radial Basis Function (RBF), which exploits gaussian activation functions in the first (or sometimes called hidden) layer. Once the inputs and the outputs have been defined, it is useful to see if the data set contains any points that violate this limits. If there are many similar examples for a given input pattern, an outlier can be classified as the one which is furthest from the median value.

Other methods that can be applied to detect outliers are the Principle Component Analysis (PCA) and Partial Least Squares (PLS). Outliers can be found by investigating points at the edges of the previously created clusters.

Liu and Gader indicated that including outlier samples in training data and using more hidden nodes than required for classification for MLP and BRF networks and proceeding an RBF with principal Component decomposition can achieve outlier rejection. The further addition of a regularization term to the PCA-RBF can achieve an outlier rejection performance equivalent or better than that of other networks without training on outliers.

The self-organizing map (SOM) is an artificial neural networks, which is trained by using unsupervised learning in order to produce a low dimensional representation of the training samples while preserving the topological properties of the input space. (Munoz & Muruzabal, 1997).

Support Vector Machine (SVM)

Support Vector Machine, introduced by V. Vapnik is a method for creating functions from a set of labelled training data. The function can be a classification function or a general regression function. In classification tasks, SVMs operate by finding a hypersurface in the space of the possible inputs, which separates the samples belonging to different classes. Such division is chosen so as to have the largest distance from the hypersurface to the nearest of the positive and negative examples.

Jordaan and Smits propose a robust model-based outlier detection approach that exploits the characteristics of the support vectors extracted by the SVM method. This method makes use of several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This approach has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are firstly removed. This allows the models in the next iteration to learn from "cleaner" data and thus reveal outliers that were masked in the initial model. The need for several iterations as well as the use of models, however, makes the on-line application of this method difficult for the not negligible computational burden. Moreover, if the data to be on-line processed come from dynamic systems, which tend to change (more or less rapidly) their conditions through time, the models update is required and this can furtherly slow the outlier detection.

Fuzzy logic

Fuzzy Logic (FL) is linked with the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which membership is a matter of degree.

Fuzzy theory is essential and is applicable to many systems — from consumer products like washing machines or refrigerators to big systems like trains or subways. Recently, fuzzy theory has been a strong tool for combining new theories (called soft computing) such as genetic algorithms or neural networks to get knowledge from real data.

Fuzzy logic is conceptually easy to understand, tolerant of imprecise data and flexible. Moreover this method can model non-linear functions of arbitrary complexity and it is based on natural language. Natural language has been shaped by thousands of years of human history to be convenient and efficient. Since fuzzy logic is built atop the structures of qualitative description used in everyday language, fuzzy logic is easy to use.

Fuzzy inference system (FIS) is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or pattern discerned. The process of fuzzy inference involves: membership functions (MF), a curve that defines how each point in the input space is mapped to a membership value or degree of membership between 0 and 1; fuzzy logic operators (and, or, not); if-then rules. Since decisions are based on the testing of all of the rules in an FIS, the rules must be combined in some manner in order to make decision. Aggregation is the process by which the fuzzy sets that represents the outputs of each rule are combined into a single fuzzy set. Aggregation only occurs once for each output variable, just prior to the final step, defuzzification.

Due to the linguistic formulation of its rule basis, the FIS provides an optimal tool to combine more criteria among those that were above illustrated according to a reasoning that is very similar to the human one. So doing, in practical application, the knowledge of the technical expert personnel can easily be exploited by the system designer.

Figure 1 depicts a scheme of the proposed method. An outlier is detected when its outlier index overcome a prefixed threshold.

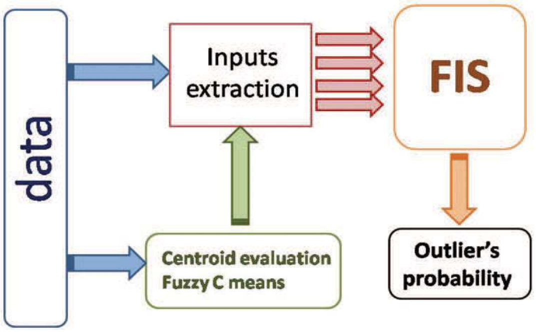

Fig.1. Block diagram depicting the overall system for outliers' detection.

Results

The proposed outlier detection method, which does not require a priori assumption on the data, has been tested in the preprocessing of data provided by a steelmaking industry, where outliers can provide indications on malfunctioning or anomalous process conditions. The method has been tested by considering that the technical personnel provided indications of those values that should be considered as outliers.

In this work two applications are proposed that use two different variables that are important to determine the quality of steel and the final destination. The variables are composed by 100 samples normalized respect their mean value.

The performance of the fuzzy logic-based method has been compared with Grubbs test and Local Outlier Factor (LOF) techniques, that are considered among the most important and widely adopted traditional outlier detection methods. The results show that the fuzzy logic based method outperforms the other approaches, but, on the other hand, the required computational time is approximately ten times greater than the time required by traditional methods, due to the increased complexity of the FIS-based evaluation.

In particular, in the first exemplar application, the considered variable represents the concentration of a chemical element extracted from the analysis made on molten steel. In this piece of database there are five outliers. The samples that are considered outliers are 3, 8, 16, 35 and 92 referring respectively to following values:

0.9349 0.9385 1.0455 0.9626 0.9541.

Figure 2 shows the result of Grubbs test. It clearly appears that only two anomalous samples have been recognized as outliers (the 3rd and 8th samples).

Figure 3 shows the result of LOF technique: only three samples are exactly classified as outliers but there are two samples that this method does not recognize as outliers.

Finally, in figure 4, is shown the result using fuzzy logic method. It clearly appears that all outliers have been correctly recognized.

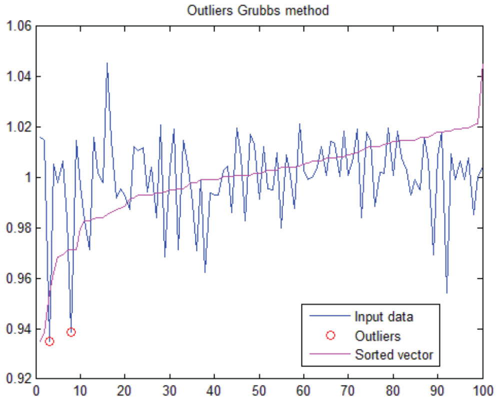

Fig. 2. First example using Grubbs method.

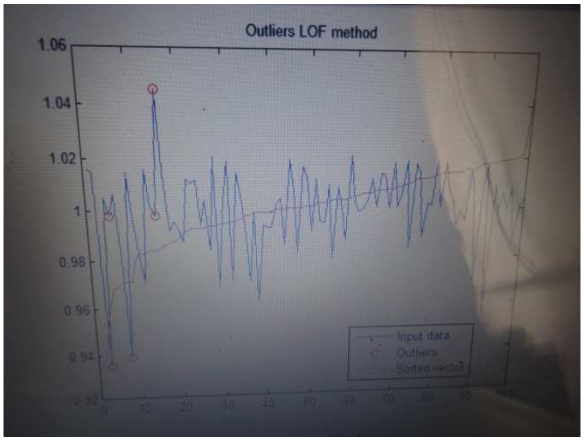

Fig. 3. First example using Local Outlier Factor method.

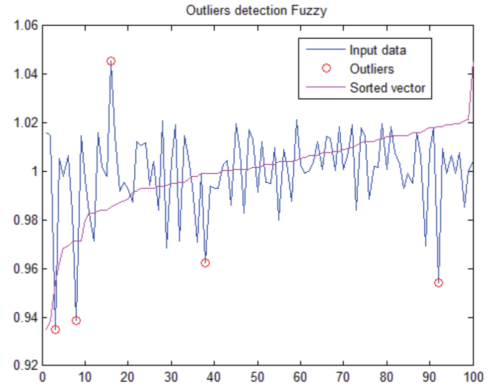

Fig. 4. First example using Fuzzy Logic method.

Conclusions and Future Work

A description of traditional approaches and of the most widely used methods within each category has been provided. As standard outlier detection methods fail to detect outliers in industrial data, the use of artificial intelligence techniques has also been proposed, because it presents the advantage of requiring poor or no a priori assumption on the considered data.

A procedure for outlier detection in a database has been proposed which exploits a Fuzzy Inference System in order to evaluate four features for a pattern that characterize its location within the database. The system has been tested on a real industrial application, where outliers can provide indications on malfunctioning or anomalous process conditions. The presented results clearly demonstrate that the Fuzzy Logic-based method outperforms the most widely adopted the traditional methods.

Future work on the FIS-based outliers detection strategy will concern the algorithm optimization in order to improve its efficiency and its on-line implementation. Moreover further tests will be performed on different applications.

References

[1] Birant, D.& Kut, A. (2006). Spatio-Temporal Detection in Large Databases, Proceedings of the 28th International Conference Information Technology Interfaces ITI 2006 June 19-22, Croatia.

[2] Bruce, A.G.; Donoho L.G.; Gao, H.Y. & Martin R.D. (2004). Denoising and robust nonlinear wavelet analysis, SPIE Proceedings Wavelet

[3] Aggarwal, C.C.; Procopiuc, C.; Wolf, J.L.; Yu, P.S. & Park, J.S. (1999). Fast algorithms for projected clustering, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 61–72, Philadephia, Pennsylvania, U.S.A.

[4] Aggarwal, R.; Gehrke, J.; Gunopulos, D. & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 94-105, Seattle, WA.

[5] Ankerst,M.; Breunig, M.M.; Kriegel, H.P. & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure, *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 49–60, June 1999, Philadelphia, Pennsylvania, U.S.A.

[6] Baldwin, J.F. (1978). Fuzzy Logic and Fuzzy Reasoning. *International Journal of Man-Machine Studies*, Vol. 11, pp. 465-480.

[7] Grubbs, F.E. (1969), Procedures for detecting outlying observations in samples, Technometrics 11, pp.1-21

[8] Hawkins, D. (1980), *Identification of Outliers*, Chapman and Hall, London.

[9] Hawkins, S.; He, X.; Williams, G.J. & Baxter, R.A. (2002). Outlier detection using replicator neural networks. *Proceedings of the 5th international conference on Knowledge Discovery and Data Warehousing.*