iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 7 **Issue:** III **Month of publication:** March 2019

**DOI:** http://doi.org/10.22214/ijraset.2019.3345

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
*Volume 7 Issue III, Mar 2019- Available at www.ijraset.com*

# A Review on Outlier Detection Approaches

Monika R. Bankar[1], Kalpana V. Metre.[2]
[1]*Student,* [2]*Professor, M.E Computer Engineering, MET's institute of engineering, Nashik, India*

***Abstract:*** ***Outlier detection is important technique in variety of domains such as digital market, detecting criminal activities, intrusion detection etc. There are 3 types of solutions: supervised, unsupervised and semi supervised for outlier detection. Supervised and unsupervised techniques require normal as well as abnormal instances in a dataset. Abnormal instances are rarely occurred and difficult to capture. A semi-supervised technique is one of the anomaly detection approach in which normal instances are provided as training. Instance can be in the form of numerical attributes, categorical attributes or can be in hybrid format. It is difficult to define distance between two categorical attributes because the values are not ordered and hence the outlier detection strategy is different for numerical and categorical attributes. In this literature work, study of different outlier detection technique is studied with their advantages and limitations. And new strategy for outlier detection is proposed.***
***Keywords: Outlier detection, distance learning, hybrid data, categorical data, semi supervised learning***

## I. INTRODUCTION

Lot of applications such as in digital marketing, detecting criminal activities, intrusion detection, fault Diagnosis, fraud detection etc. make use of outlier detection technique. The outlier is the instance identification process which has abnormal behaviour than regular instances. Outliers are unexpected, extreme instances in a dataset. Outliers are also called as anomalies. Outlier detection and noise removal are two closely related topics in data mining domain. Noisy data creates obstruction in data analysis and hence such noisy data removal is pre-processing task before applying any data mining technique such as clustering, classification, regression, etc. The outlier detection is the main stream multidisciplinary research area. Outlier detection and removal technique removes abnormal instances from data. The normal instances follow some data distribution and occurrence strategy. The outlier detection can also be called as binary class classification problem in which incoming instances are classified in normal or anomalous class. The classification technique is mainly categorized in two sections based on the type of data and availability of training data. The data may contain numeric information or categorical information. Based on the availability of training data the classification technique is classified in 3 different approaches.

1) *Supervised:* In supervised learning, training data is available. The training data contains binary class information. For outlier detection abnormal class is having very few instances. Such data is called as imbalanced data. The imbalanced data has poor classification result due to lack of sufficient training information.
2) *Unsupervised:* In unsupervised learning technique, no training data is available. The technique works based on the data density or distance technique. The abnormal instances are separated from normal data by some geometrical distance from normal data instances. The unsupervised technique uses clustering algorithm for grouping instances. The normal instances are closer to each other than outliers.
3) *Semi-Supervised:* In semi supervised approach only single class data is available for training. Generally normal instances are provided in the training phase.

Lot of research work has been done in supervised and unsupervised outlier detection technique. For supervised learning, instances of normal and abnormal class are required. The abnormal instances are rarely occurred and highly expensive. The collection of balanced dataset for supervised outlier detection technique is infeasible. To overcome this problem semi supervised approach provides better solution. Semi supervised approach do not require abnormal instances for training. In training phase, model is proposed based normal instances. In testing phase, instances are classified in abnormal class if they are far away from normal instances. The second scenario of classification technique is type of data. In many applications data is present with numeric attributes. Each instance is described with m dimensional feature vector space. The attributes with numeric values have wide range of possible proximity measures. Lot of research has been done on outlier detection technique with numerical attributes.

Lot of applications generates categorical data. In Categorical data, attributes represents an unordered nominal values. These unordered nominal values cannot be mapped to the numerical values without loss of information. For example user marital status, profession, job profile, education has one value from set of possible values. User cannot map these values to numeric values directly and hence cannot find the distance between these attribute values directly. In the following section various outlier detection and removal techniques are discussed followed by problem Formulation.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
*Volume 7 Issue III, Mar 2019- Available at www.ijraset.com*

## II. LITERATURE REVIEW

### *A. Unsupervised And Semi Supervised Outlier Detection*

Local outlier detection (LOF) is an unsupervised outlier detection technique. This is distance based outlier detection technique. knn is used to finds k neighbor of each point. A point having very few nearest neighbor are outlier points. This technique is depending on the parameter k and works on numerical dataset[2].

Unlike distance based technique, cluster based outlier detection is again an unsupervised technique. C. H. Wang introduces a kernel based clustering technique for outlier detection. A market basket analysis is done in this technique and outlier customer market records are extracted. Kriegel et al. proposes a technique of angle based outlier detection. This is parameter free approach. It uses variance of angles between data points[3][4].

The above methods focuses on numerical dataset and not be applicable for categorical dataset. The numerical method measures cannot be directly applied to the categorical dataset. Direct application of such measure may negatively affect the result.

Partial training data is required in semi-supervised approach. Most of the work on semi supervised outlier detection technique is done in last 15 years. Support vector machine based semi supervised approach is proposed first using one class SVM classification problem. This technique maps input data to the high dimensional feature vector space. This technique follows iterative algorithm to find maximal margin hyperplanes that covers the training data.[5]

Least-square importance fitting is a statistical outlier detection technique. It calculates the 'importance factor' of each training instance. Test instances are compared with training instances. Outlier instances in test set are identified using density ratio. If the density ratio of training and test set is high then such instances are identified as outliers.[6]

Feature regression and classification FRaC is a semi supervised approach. Normal instances are used to build classification model. This technique compares the test instances with the feature model and instances which are not following the model rules are treated as anomalous. FRaC is not designed for categorical data.[7]

### *B. Outlier Detection over Categorical Data*

For categorical dataset outliers are identified using frequent itemset extraction technique. Each item is nothing but a categorical value. Items those are not frequently occurred are treated as outliers.[8]

A fast greedy algorithm is proposed by Z. He, S. Deng, X. Xu. This greedy algorithm deals outliers detection as optimization problem. It uses maximal entropy technique. Koufakou et al. proposes a score based evaluation. The score is nothing but an occurrence frequency. This technique assigns score to each attribute value pair in a dataset. The instances having lowest score for all attribute values are considered as outliers. Both techniques are unsupervised and need threshold value for outlier detection. [9][10]

Akoglu et al proposes a technique based on minimum description length (MDL) principle. The proposed technique uses pattern-based compression mechanism. Instances having higher compression cost are treated as outliers.[11]

Smets and Vreeken proposes a semi supervised technique for outlier detection over categorical data based on compression technique. The technique initially uses frequent itemset extraction with minimum support value as an input. This technique is time consuming and its accuracy is less as compared to Akoglu et al technique[12].

D. Ienco, R. G. Pensa, and R. Meo proposes a semi supervised technique for outlier detection over categorical dataset. It uses distance learning technique to characterize the normal class as a partial training set in semi supervised mode. Test instances are compared with normal class training data. The outlier score of each instance is calculated based on DILCA algorithm. Based on the outlier score top k outliers are extracted.[1]

## III.PROPOSED METHODOLOGY

The classification technique is mainly categorized in two sections based on:

1) *Type Of Data*
2) *Availability Of Training Data*

Most of the existing techniques find outliers using supervised or unsupervised method. The supervised or unsupervised method requires normal as well as abnormal instances for training. The abnormal instances are rarely occurred and collecting data for abnormal instances is quite expensive and difficult task. To find outliers by only providing one domain training set is required. Semi supervised technique requires only one type of data for training. Most of the application generates hybrid data contains numeric as well as categorical attributes. A semi supervised technique is required for outlier detection over such hybrid dataset.

### *A. Preliminaries*

*1) Symmetric Uncertainty:* This is a Non-linear estimation. It defines the correlation between two attributes X and Y as:

$$SU(X,Y) = \frac{2 * gain(X|Y)}{H(X)+H(Y)}$$

Where gain is defined as:

Gain(X|Y) = H(X) - H(X|Y)

H(X) is the entropy of a discrete random attribute X,

H(X|Y ) is the conditional entropy.

The value of SU lies between 0 to 1. If the SU value is 1 then the two attributes are highly correlated. If the value is 0 then two attributes are independent.

*2) Context-Based Distance:* Content based distance between value pair Yi and Yj of target attribute is calculated using probability distribution with each context attribute X. It can be calculated as:

$$D(yi,yj) = \sqrt{\frac{\sum_{x\in C(Y)} \sum_{xk\in X} (P(yi|xk) - P(yj|xk))^2}{\sum_{x\in C(Y)} |X|}}$$

*3) DILCA Distance:* The distance between two instances d1 and d2 with categorical values is defined as:

$$Dist(d1,d2) = \sqrt{\sum_{MXi\, \varepsilon M} mi(d1[Xi], d2[Xi])^2}$$

Where d1[Xi] and d2[Xi] are values of attribute X for instance d1 and d2 respectively.

M is the learning model and MXi is the matrix for attribute X. The value of matrix is the context based distance between values Xi and Xj [6]

*4) Outlier Score:* The outlier score of instance dp from test data is computed as

$$OS(t) = \sum_{p-1}^{k} dist(t, dp)$$

Where t Ɛ Training data instance and dist is the DILCA distance.

### *B. Proposed System*

Following figure shows the architecture of system.

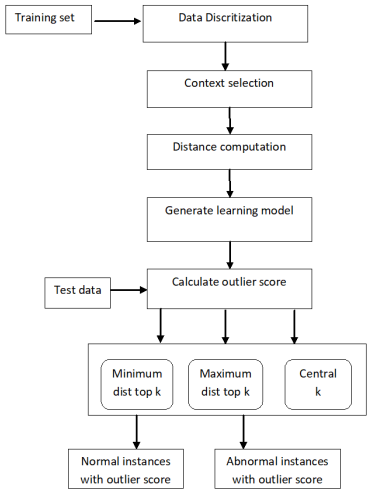



Figure 1: System Architecture

*1) System Flow:* The system is executed in two phases. The first phase is training phase and the other one is testing phase.

The system is executed in two phases. The first phase is training phase and the other one is testing phase.

*a)* For hybrid dataset, numerical attributes are converted in to categorical values using discritization process. The data discritization uses equal binning process. Then system ask user to select target attribute for context based distance learning. Based on the target attribute learning model matrices M={MXi} are generated for other attributes Xi in the dataset. The value of matrices represents the context based distance of two attribute values Xi. This distance is calculated using training data.

*b)* Based on the learning model matrices, the distance of test instance is evaluated with training data using DILCA distance formula.[13] Using DILCA distance of test instance with every instance of training data, the outlier score is evaluated.

The outlier score of each instance is calculated and following results are generated:

*i)* Minimum distance top k: For test instance t, top k nearest instances from training data are enlisted.

*ii)* Maximum Distance Top-k : For test instance t, top k instances from training data are enlisted which are far away from test instance t.

*iii)* Central k: This is the summary of training data. The outlier score of each training data instance is evaluated. The most central instances of training data are identified. The central instances have minimum outlier score.

## IV. CONCLUSIONS

Most of the existing technique works on numerical dataset for outlier detection. The detection techniques are classified as supervised, unsupervised and semi supervised based on availability of training data. The outlier detection is treated as binary class classification problem. For training anomalous instances data is rarely occurred hence semi supervised is better solution for outlier detection. Most system generates numerical as well as categorical data. A system is required for outlier detection over hybrid dataset. A new system is proposed with semi-supervised approach for outlier detection system. The system works on hybrid dataset which contains numerical as well as categorical attributes. The data discritization process converts the numerical attributes to categorical format. For categorical data system uses context-based distance learning. The distance is evaluated using attributes values distribution over data objects. Based on the training data outlier score of test data is calculated. Using outlier score dataset summary is generated by finding minimum top k, maximum top k and central outliers.

## REFERENCES

[1] Dino Ienco, Ruggero G. Pensa, and Rosa Meo, "A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data," in IEEE Transactions on Neural Networks and Learning Systems, Vol. 28 , no. 5 , pp. 1017 - 1029, 2017

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in Proc. ACM SIGMOD Conf. Manage. Data, Dallas, TX, USA, May 2000, pp. 93–104.

[3] C.-H. Wang, “Outlier identification and market segmentation using kernel-based clustering techniques,” Expert Syst. Appl., vol. 36, no. 2, pp. 3744–3750, 2009.

[4] H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in Proc. 14th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Las Vegas, NV, USA, Aug. 2008, pp. 444–452.

[5] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” Neural Comput., vol. 13, no. 7, pp. 1443–1471, 2001.

[6] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, “Statistical outlier detection using direct density ratio estimation,” Knowl. Inf. Syst., vol. 26, no. 2, pp. 309–336, 2011.

[7] K. Noto, C. Brodley, and D. Slonim, “FRaC: A feature-modeling approach for semi-supervised and unsupervised anomaly detection,” Data Mining Knowl. Discovery, vol. 25, no. 1, pp. 109–133, 2012.

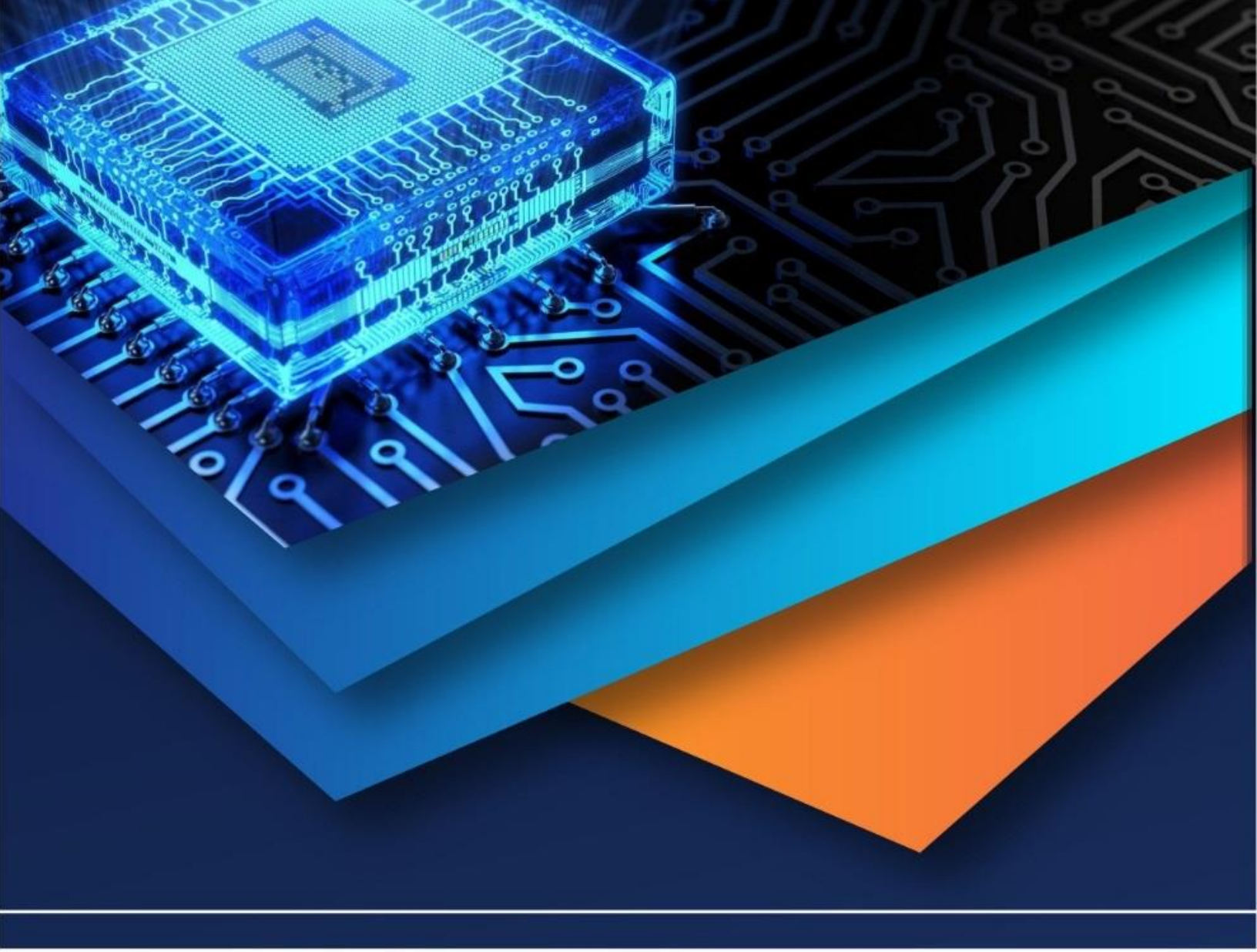
[8] Z. He, X. Xu, J. Z. Huang, and S. Deng, “FP-outlier: Frequent pattern based outlier detection,” Comput. Sci. Inf. Syst., vol. 2, no. 1, pp. 103–118, 2005.

[9] Z. He, S. Deng, X. Xu, and J. Z. Huang, “A fast greedy algorithm for outlier mining,” in Proc. 10th PAKDD, Singapore, Apr. 2006, pp. 567–576.

[10] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos, “Fast and reliable anomaly detection in categorical data,” in Proc. 21st ACM CIKM, Maui, HI, USA, Oct. 2012, pp. 415–424.

[11] K. Smets and J. Vreeken, “The odd one out: Identifying and characterising anomalies,” in Proc. SIAM Int. Conf. Data Mining, Mesa, AZ, USA, Aug. 2011, pp. 804–815.

[12] D. Ienco, R. G. Pensa, and R. Meo, “From context to distance: Learning dissimilarity for categorical data clustering,” ACM Trans. Knowl. Discovery Data, vol. 6, no. 1, 2012, Art. ID 1.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY