3rd International Conference on Recent Innovations in Science Engineering and Management
Sri Venkateswara College of Engineering and Technology,Srikakulam,Andhra Pradesh (ICRISEM-16)
27 February 2016, www.conferenceworld.in ISBN: 978-81-932074-1-3

OUTLIER DETECTION USING WEIGHTED HOLOENTROPY

Ms. Manasi V. Harshe

Computer Engineering Department, TSSM's BSCOER, Narhe, Pune,, Maharashtra, (India)

ABSTRACT

Outlier detection can usually be considered as a pre-processing step for locating, in a data set, those objects that do not conform to well-defined notions of expected behavior. It is very important in data mining for discovering novel or rare events, anomalies, vicious actions, exceptional phenomena, etc. Several methods are investigated for outlier detection corresponding to categorical data sets. This problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data. In the previous work, holoentropy was used for outlier detection, as the weightage function is based on the reverse sigmoid function. In the proposed method, logistic sigmoid function related to hyperbolic tangent will be used as weightage function for finding the outlier data point. The advantage of this weightage function is that it can differentiate or distribute the outlier data points effectively as compared with the reverse sigmoid function. The method is implemented with four phases. In the first phase, the data is read out through programming and dynamic entropy computation is done in the second phase which consists of data points extraction, probability computation and dynamic entropy computation using logistic sigmoid function related to hyperbolic tangent. In the third phase, dynamic entropy related to all the data points are sorted out and the top N point are selected as outlier data point and finally the accuracy is computed for evaluating the proposed method whether the outlier data points are detected correctly.

Keywords*: *Holoentropy, outlier detection, weighted function.

1. INTRODUCTION

Of all the data mining techniques that are in vogue, outlier detection comes closest to the metaphor of mining for nuggets of information in real-world data. It is concerned with discovering the exceptional behavior of certain objects. An outlier is defined as a data point which is very different from the rest of the data based on some measure. Such a point often contains useful information on abnormal behavior of the system described by the data. Outliers are a well-known problem in all experimental scientific and industrial fields. The outlier detection are classified according to the availability of labels in the training data sets, there are three broad categories: supervised, semi-supervised, and unsupervised approaches. In principle, models within the supervised or the semi-supervised approaches all need to be trained before use, while models adopting the unsupervised approach do not include the training phase. Moreover, in a supervised approach a training set should be provided with labels for anomalies as well as labels of normal objects, in contrast with the training set with normal object labels alone required by the semi-supervised approach. On the other hand, the unsupervised

approach does not require any object label information. Thus the three approaches have different prerequisites and limitations, and they fit different kinds of data sets with different amounts of label information.

The supervised anomaly detection approach learns a classifier using labeled objects belonging to the normal and anomaly classes, and assigns appropriate labels to test objects. The semi-supervised anomaly detection approach primarily learns a model representing normal behavior from a given training data set of normal objects, and then calculates the likelihood of a test object's being generated by the learned model. The semi-supervised outlier detection methods could perform better than unsupervised methods thanks to additional label information, but such outlier samples for training are not always available in practice. Furthermore, the type of outliers may be diverse and thus the semi-supervised methods—learning from known types of outliers—are not necessarily useful in detecting unknown types of outliers. The unsupervised anomaly detection approach detects anomalies in an unlabeled data set under the assumption that the majority of the objects in the data set are normal. Among the unsupervised approaches, distance-based methods distinguish an object as outlier on the basis of the distances to its nearest neighbors. These approaches differ in the way the distance measure is defined, but in general, given a data set of objects, an object can be associated with a weight or score, which is, intuitively, a function of its k-nearest neighbors distances quantifying the dissimilarity of the object from its neighbours. Identification of potential outliers is important for the following reasons.

- An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).
- In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation.

Interestingly, the difference in weighted holoentropy can be estimated, especially when only one object is removed, without having to estimate attribute probabilities. This opens up the possibility of an efficient heuristic approach to solving optimization problem. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution.

The proposed holoentropy assigns equal importance to all the attributes, whereas in real applications, different attributes often contribute differently to form the overall structure of the data set. The proposed weighting method computes the weights directly from the data whereas we introduce biased weights on holoentropy calculation so that the nonlinearities in the data can be handled efficiently and is motivated by increased effectiveness in practical applications rather than by theoretical necessity. Existing systems are more dependable on user defined parameters and very few methods are dealing with unsupervised categorical data. So there should be appropriate method which will able to deal large scale data without requirement of any user defined parameter.

II. RELATED WORK

2.1 Proximity-based Techniques

Proximity-based techniques are simple to implement and make no prior assumptions about the data distribution model. They are suitable for both type 1 and type 2 outlier detection. However, they suffer exponential

computational growth as they are founded on the calculation of the distances between all records. The computational complexity is directly proportional to both the dimensionality of the data m and the number of records n. Hence, methods such as k-nearest neighbour (also known as instance-based learning and described next) with O(n2m) runtime are not feasible for high dimensionality data sets unless the running time can be improved. There are various flavours of k-Nearest Neighbour (k-NN) algorithm for outlier detection but all calculate the nearest neighbours of a record using a suitable distance calculation metric such as Euclidean distance or Mahalanobis distance.

Ramaswamy et al. [22] introduce an optimised kNN to produce a ranked list of potential outliers. A point p is an outlier if no more than n − 1 other points in the data set have a higher Dm (distance to mth neighbour) where m is a user-specified parameter. V is most isolated followed by X, W, Y then Z so the outlier rank would be V, X, W, Y, Z. This approach is susceptible to the computational growth as the entire distance matrix must be calculated for all points (ALL k-NN) so Ramaswamy et al. include techniques for speeding the k-NN algorithm such as partitioning the data into cells. If any cell and its directly adjacent neighbours contain more than k points, then the points in the cell are deemed to lie in a dense area of the distribution so the points contained are unlikely to be outliers. If the number of points lying in cells more than a pre-specified distance apart is less than k then all points in the cell are labelled as outliers. Hence, only a small number of cells not previously labelled need to be processed and only a relatively small number of distances need to be calculated for outlier detection. Authors have also improved the running speed of k-NN by creating an efficient index using a computationally efficient indexing structure [23] with linear running time.

Knorr & Ng [24] introduce an efficient type 1 kNN approach. If m of the k nearest neighbours (where m<k) lie within a specific distance threshold d then the exemplar is deemed to lie in a sufficiently dense region of the data distribution to be classified as normal. However, if there are less than m neighbours inside the distance threshold then the exemplar is an outlier. A very similar type 1 approach for identifying land mines from satellite ground images [25] is to take the m-th neighbour and find the distance Dm. If this distance is less than a threshold d then the exemplar lies in a sufficiently dense region of the data distribution and is classified as normal. However, if the distance is more than the threshold v alue then the exemplar must lie in a locally sparse area and is an outlier. This has reduced the number of data-specific parameters from Knorr & Ng's [24] have classified outliers relatively distant from their neighbours. This is less susceptible to the computational growth than the ALL k-NN approach as only the k nearest neighbours needs to be calculated for a new exemplar rather than the entire distance matrix for all points.

2.2 Parametric Methods

Many of the methods described do not scale well unless modifications and optimisations are made to the standard algorithm. Parametric methods allow the model to be evaluated very rapidly for new instances and are suitable for large data sets; the model grows only with model complexity not data size. However, they limit their applicability by enforcing a pre-selected distribution model to fit the data. If the user knows their data fits such a distribution model then these approaches are highly accurate but many data sets do not fit one particular model. One such approach is Minimum Volume Ellipsoid estimation (MVE) [28] which fits the smallest permissible ellipsoid volume around the majority of the data distribution model (generally covering 50% of the

data points). A similar approach, Convex Peeling peels away the records on the boundaries of the data distribution's convex hull and thus peels away the outliers. In contrast MVE maintains all points and defines a boundary around the majority of points. In convex peeling, each point is assigned a depth. The outliers will have the lowest depth thus placing them on the boundary of the convex hull and are shed from the distribution model.

Both MVE and Convex Peeling are robust classifiers that fit boundaries around specific percentages of the data irrespective of the sparseness of the outlying regions and hence outlying data points do not skew the boundary. Both however, rely on a good spread of the data. Both MVE and convex peeling are only applicable for lower dimensional data sets [29] (usually three dimensional or less for convex peeling) as they suffer the Curse of Dimensionality where the convex hull is stretched as more dimensions are added and the surface becomes too difficult to discern. Torr and Murray [30] also peel away outlying points by iteratively pruning and re-fitting. Faloutsos et al. [21] recommend retaining sufficient components so the sum of the eigenvalues of all retained components is at least 85% of the sum of all eigenvalues. They use the principal components to predict attribute values in records by finding the intersection between the given values for the record (i.e., excluding the omitted attribute) and the principal components. If the actual value for an attribute and the predicted value differ then the record is flagged as an outlier.

2.3 Non-Parametric Methods

Many Statistical methods described in this section have data-specific parameters ranging from the k values of k-NN and k-means to distance thresholds for the proximity-based approaches to complex model parameters. Other techniques such as those based around convex hulls and regression and the PCA approaches assume the data follows a specific model. These all require apriori data knowledge. Such information is often not available or is expensive to compute. Many data sets simply do not follow one specific distribution model and are often randomly distributed. Hence, these approaches may be applicable for an outlier detector where all data is accumulated beforehand and may be pre-processed to determine parameter settings or for data where the distribution model is known. Non-parametric approaches, in contrast are more flexible and autonomous.

Statistical approaches were the earliest algorithms used for outlier detection. Some of the earliest are applicable only for single dimensional data sets. In fact, many of the techniques described in both [13] and [14] are single dimensional or at best univariate. One such single dimensional method is Grubbs' method (Extreme Studentized Deviate) [15] 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 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 user parameters as all parameters are derived directly from data. However, the technique is susceptible to the number of exemplars in the data set. The higher the number of records the more statistically representative the sample is likely to be. Statistical models are generally suited to quantitative real-valued data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical (numerical) processing. This limits their applicability and increases the processing time if complex data transformations are necessary before processing.

Probably one of the simplest statistical outlier detection techniques described here, Laurikkala et al. [16] use informal box plots to pinpoint outliers in both univariate and multivariate data sets. This produces a graphical representation and allows a human auditor to visually pinpoint the outlying points. Their approach can handle real-valued, ordinal and categorical (no order) attributes. Box plots plot the lower extreme, lower quartile, median, upper quartile and upper extreme points. The outliers are the points beyond the lower and upper extreme values of the box plot. Laurikkala et al. suggest a heuristic of 1.5 × inter-quartile range beyond the upper and lower extremes for outliers but this would need to vary across different data sets. For multivariate data sets the authors note that there are no unambiguous total orderings but recommend using the reduced sub-ordering based on the generalised distance metric using the Mahalanobis distance measure. The Mahalanobis distance measure includes the inter-attribute dependencies so the system can compare attribute combinations. The authors found the approach most accurate for multivariate data where a panel of experts agreed with the outliers detected by the system. For univariate data, outliers are more subjective and may be naturally occurring, for example the heights of adult humans, so there was generally more disagreement. Box plots make no assumptions about the data distribution model but are reliant on a human to note the extreme points plotted on the box plot.

Statistical models use different approaches to overcome the problem of increasing dimensionality which both increases the processing time and distorts the data distribution by spreading the convex hull. Some methods preselect key exemplars to reduce the processing time [17, 18]. As the dimensionality increases, the data points are spread through a larger volume and become less dense. This makes the convex hull harder to discern and is known as the "Curse of Dimensionality". The most efficient statistical techniques automatically focus on the salient attributes and are able to process the higher number of dimensions in tractable time. However, many techniques such as k-NN, neural networks, Minimum Volume Ellipsoid or Convex Peeling described in this survey are susceptible to the Curse of Dimensionality. These approaches may utilise a preprocessing algorithm to preselect the salient attributes [18, 19]. These feature selection techniques essentially remove noise from the data distribution and focus the main cluster of normal data points while isolating the outliers. Only a few attributes usually contribute to the deviation of an outlier case from a normal case. An alternative technique is to use an algorithm to project the data onto a lower dimensional subspace to compact the convex hull [20] or use Principal Component Analysis [21].

III. IMPLEMENTATION DETAILS

The proposed methodology for outlier detection is explained in this section. In the previous work, holoentropy was used for outlier detection as the weightage function is based on the reverse sigmoid function. In the proposed method, logistic sigmoid function related to hyperbolic tangent will be used as weightage function for finding the outlier data point. The advantage of this weightage function is that it can differentiate or distribute the outlier data points effectively as compared with the reverse sigmoid function.The block diagram of the proposed methodology is explained in figure 2.

The method is implemented with four phases.

- In the first phase, the data is read out through programming and dynamic entropy computation is done.

- In the second phase which consists of data points extraction, probability computation and dynamic entropy computation using logistic sigmoid function related to hyperbolic tangent.
- In the third phase, dynamic entropy related to all the data points are sorted out and the top N point are selected as outlier data point.
- Finally in fourth phase the accuracy is computed for evaluating the proposed method whether the outlier data points are detected correctly.

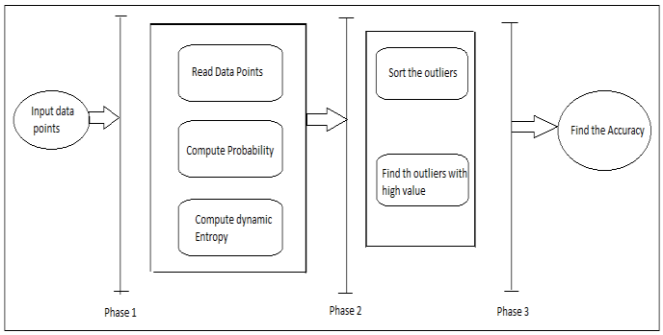

Fig.1 Block diagram of the proposed outlier detection method

IV. PROPOSED ALGORITHM

Step 1: Take the input in the form of array for dataset D

Step 2: Calculate the frequency of data samples as outliers

Step 3: Find the unique values of data sample attributes

Step 4: Calculate the Dynamic Entropy using this frequency and unique values

Step 5: Calculate the Holoentropy of existing algorithm

Step 6: Calculate frequency using

Frequency = (current frequency value)/(total input size)

Step 7: Calculate entropy for proposed algorithm using the above Frequency

Step 8: Calculate entropy using weighing factor and hyperbolic tangent sigmoid function.

Step 9: Calculate Holoentropy using above values

Step 10: Sort the outliers with highest entropy values.

Step 11: Find the accuracy of the respective outlier data.

4.1 Proposed Mathematical Model

The weighted factor is computed based on log-sigmoid function in the in the exiting work. A sigmoid function is a mathematical function having an "S" shape (sigmoid curve). Often, sigmoid function refers to the special case of the logistic function. The logistic sigmoid function is often interpreted as probabilities (in, say, logistic regression).

The log-sigmoid function is defined as,

$$L = \frac{1}{1+\exp\left(-\left(DEi_{j)}\right)\right)} \quad \text{- (1)}$$

The hyperbolic tangent sigmoid function defined in the literature is given as follows,

$$T = \frac{2}{(1+\exp(-2*DEi_j))-1} \quad \text{- (2)}$$

The *tanh* function, a.k.a. hyperbolic tangent function, is a rescaling of the logistic sigmoid. The simple sigmoids, defined to be odd, asymptotically bounded, completely monotone functions in one variable, and the Hyperbolic sigmoids, a proper subset of simple sigmoids and a natural generalization of the hyperbolic tangent. The class of hyperbolic sigmoids includes a surprising number of well known sigmoids. The regular structure of the simple sigmoids often makes a theory tractable, paving the way for more general analysis.

In the proposed method, hyperbolic tangent sigmoid function is taken and it is included in the weight factor computation as like,

$$WF_{ij} = 2\left(1 - \frac{2}{(1+\exp(-2*DEi_j))-1}\right) \quad \text{-(3)}$$

Step 1: Dynamic entropy computation

The entropy can be used as a global measure in outlier detection. In information theory, entropy means uncertainty relative to a random variable: if the value of an attribute is unknown, the entropy of this attribute indicates how much information we need to predict the correct value. A subset of objects is good outlier candidates if their removal from the data set causes significant decrease of the entropy of the data set.

The objective of this work is to find the dynamic entropy of every data point excepting the current data point. This procedure will give the dynamic entropy for every data point. In order to find the dynamic entropy of data point d_i, DE_{ij} is found out and then, average is computed.

$$DE_i = \frac{1}{m}\sum_{j=1}^{m} DE_{ij}$$

Where, DE_{ij} is the dynamic entropy belonging to the attribute of a_j for the i the data point. This computation requires

the unique data symbols available in j the attribute. The final dynamic entropy of data point d_i is computed as follows,

$$FDE_{D\setminus\{d_i\}} = DE_{ij} * WF_{ij}$$

The weighted factor is computed based on log-sigmoid function in the exiting work. But, entropy alone is not a good enough measure for outlier detection and the contribution of the total correlation is necessary. The holoentropy is defined as the sum of the entropy and the total correlation of the random vector Y, and can be expressed by the sum of the entropies on all attributes. The holoentropy assigns equal importance to all the attributes, whereas in real applications, different attributes often contribute differently to form the overall structure of the data set. Then, a simple method for weighting attributes and then modify the holoentropy by incorporating the attribute weights is given. The proposed weighting method computes the weights directly from the data and is motivated by increased effectiveness in practical applications rather than by theoretical necessity. The mathematical formulation is given in section 4.4.

Step 2. Outlier detection

The definition of detecting outliers is based on the weighted holoentropy, supposing that the number of the desired outliers 'o' is given. A set of candidates is the best if its exclusion from the original data set X causes the greatest decrease in the weighted holoentropy value, compared to all the other subsets of X. This set is taken as final outlier from the input data and it is removed from the database

NP-Hard analysis:

In the proposed algorithm, the attribute weights, the holo-entropy of all the objects and the sort search to find the top-outlier candidates are computed. The time complexity of computing holoentropy and outlier data points is O(mn), and the time cost of top-outlier searching is O(nlog(No)). Since the value of log(No) is always much smaller than the number of attributes for real applications, the final time complexity of the proposed algorithm can be written as O(nm).

V. PREREQUISITES

This section presents the experimental results and its discussion of the proposed hyperbolic tangent-sigmoid weighted holoentropy–based outlier detection method. Experimental set up and description about the datasets taken for experimentation is explained in the following section. Along with, experimental results and analysis has been performed to ensure the performance of the proposed hyperbolic tangent-sigmoid weighted holoentropy–based outlier detection method.

Dataset description : The proposed system is experimented with the two different datasets namely, Clevland and Iris data.

Cleveland data: The data used in this study is the Cleveland Clinic Foundation Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Iris data*:* The data set contains three categories of 50 objects each, where each category refers to a type of iris plant. Here, third category is taken as outlier data point.

5.1 Evaluation Matrics

Evaluation metric is important for any outlier detection method to evaluate the performance. Here, the performance is completely analyzed with metric given as below:

$$Accuracy = \frac{Correctly \ \text{detected outlier} \ samples}{Total \ samples}$$

5.2 Screenshots

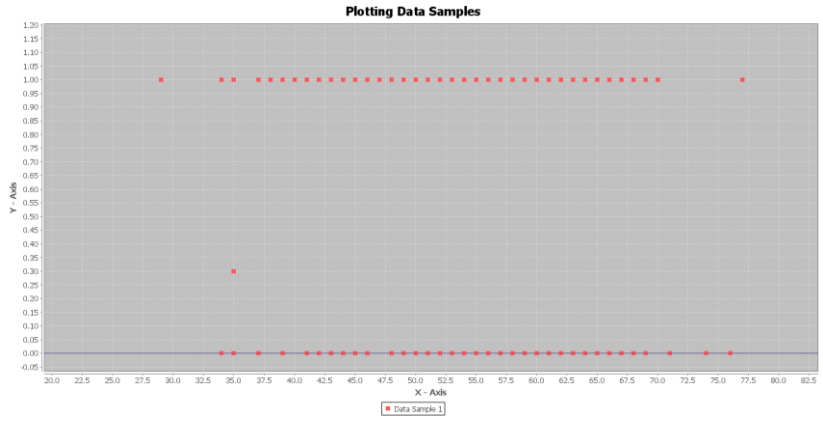

Fig. 2 GUI of the proposed method for Cleveland Dataset input

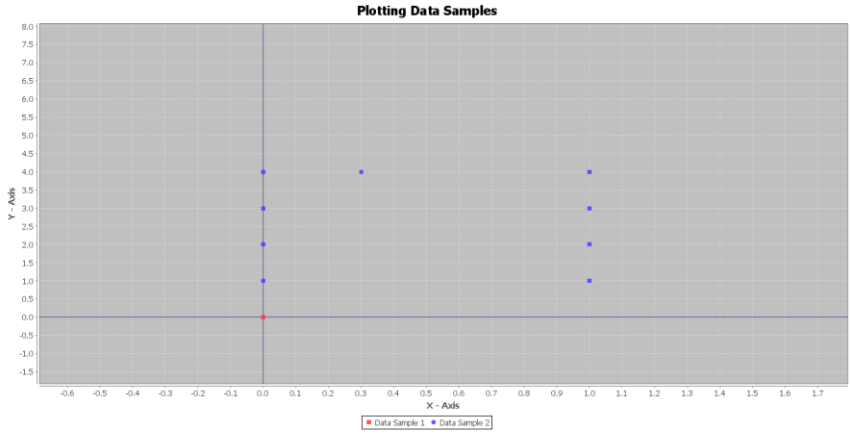

Fig. 3 GUI of the proposed method for Cleveland Dataset output with outliers

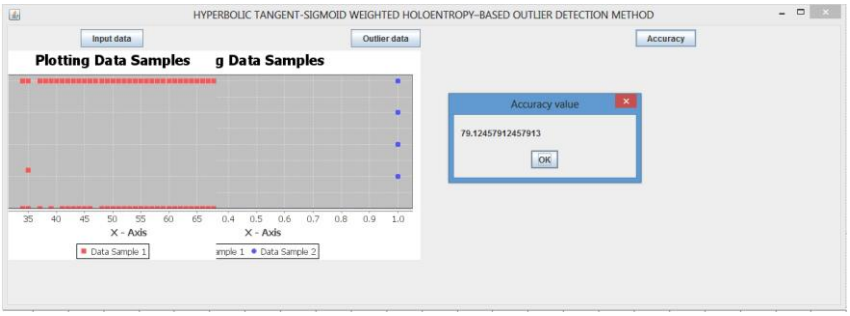

Fig.4 Outlier Accuracy of Cleveland Dataset

5.3 Data table discussion

Sr. No.	No. Of Outliers	Existing Method	Proposed Method
1.	15	77.44	79.12
2.	22	77.44	79.46
3.	27	77.44	79.46
4.	36	77.77	79.46

Table 1. Cleveland data Accuracy Percentage of existing & proposed methods

VI. APPLICATIONS

Since outlier detection is useful in various applications it has been a active research topic in statistics, machine learning, and data mining communities for decades. Some of the application areas are elaborated below.

- **Intrusion Detection :** Intrusion detection refers to detection of malicious activity(break-ins, penetrations, and other forms of computer abuse) in a computer related system interesting from a computer security perspective. Being different from normal system behaviour, intrusion detection is a perfect candidate for applying outlier detection techniques. The key challenges for outlier detection are :-
  - **Huge Data Volume:** This calls for computationally efficient techniques.
  - **Streaming Data:** This requires on-line analysis.
  - **False alarm rate:** Smallest percentage of false alarms among millions of data objects can make be overwhelming for an analyst.
  - **Labeled data not usually available for Intrusions:** This gives preference to semisupervised and unsupervised outlier detection techniques.
- **Fraud Detection :** Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market etc. Malicious users could be actual customers of the organization or resorting to identity theft (posing as customers). The detection activity aims at detection of unauthorized consumption of resources provided by the organization to prevent economic losses.
- **Credit Card Fraud Detection:** Outlier detection techniques are applied to detect :-

Fraudulent Applications for Credit Card: This is similar to detecting insurance fraud

Fraudulent Usage of Credit Card: Associated with credit card thefts

- **Mobile Phone Fraud Detection :** In this activity monitoring problem the calling behavior of each account is scanned to issue an alarm when an account appears to have been misused.
- **Insurance Claim Fraud Detection :** An important problem in the property-casualty insurance industry is claims fraud, e.g. automobile insurance fraud. Individuals and conspiratorial rings of claimants and providers manipulate the claim processing system for unauthorized and illegal claims.
- **Insider Trading Detection :** Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on (or leaking) inside information before the information is made public.
- **Medical and Public Health Outlier Detection :** The data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. The data might also have temporal as well as spatial aspect to it. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors.
- **Industrial Damage Detection :** Industrial units suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is usually sensor data recorded using different sensors and collected for analysis.

- **Image Processing :** Outlier detection here aims to detect changes in an image over time (motion detection) or in regions which appear abnormal on the static image.
- **Outlier Detection in Text Data :** Outlier detection techniques in this domain primarily detect novel topics or events or news stories in a collection of documents or news articles. The outliers are caused due to a new interesting event or an anomalous topic. The data in this domain is typically high dimensional and very sparse. The data also has a temporal aspect since the documents are collected over time.

VII. CONCLUSION

Outlier detection encompasses aspects of a broad spectrum of techniques. Many techniques employed for detecting outliers are fundamentally identical but with different names chosen by the authors. For example, authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection or exception mining. The objective of this work is to find the dynamic entropy of every data point excepting the current data point. In the proposed method, logistic sigmoid function related to hyperbolic tangent is used as weightage function for finding the outlier data point. The advantage of this weightage function is that it can differentiate or distribute the outlier data points effectively as compared with the reverse sigmoid function. The proposed weighting method computes the weights directly from the data and is motivated by increased effectiveness in practical applications rather than by theoretical necessity.

The future work can be in the direction of complexity free algorithm for finding outlier data point rather than finding entropy for a data point with respect to whole database. Also, apart from entropy-based computation, gini index, information gain can be effectively applied for finding the outlier data points.

VIII. ACKNOWLEDGMENT

The author would like to thank the researchers as well as the publishers for making their resources available and teachers for their guidance. Also thankful to reviewer for their valuable suggestions. I also thank the college authorities for providing the required infrastructure and support. Finally, I would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori, *"Inlier-based Outlier Detection via Direct Density Ratio Estimation"*, Eighth IEEE International Conference on Data Mining, 2008.

[2] Peter Filzmoser, Karel Hron, *"Outlier Detection for Compositional Data Using Robust Methods"*, Math Geosci, vol. 40, pp. 233–248, 2008.

[3] Yixin Chen, Xin Dang, Hanxiang Peng, Henry L. Bart, *"Outlier Detection with the Kernelized Spatial Depth Function"*, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, February 2009.

[4] Tao Chen, Elaine Martin, Gary Montague, *"Robust probabilistic PCA with missing data and contribution analysis for outlier detection"*, Computational Statistics & Data Analysis, Vol. 53, No. 10, August 2009.

[5] Guido Buzzi-Ferraris, Flavio Manenti, *"Outlier detection in large data sets"*, Computers and Chemical Engineering, vol. 35, pp. 388–390, 2011.

[6] Jing Liu, HuiFang Deng, *"Outlier detection on uncertain data based on local information"*, Knowledge-Based Systems, vol. 51, pp. 60–71, 2013.

[7] Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori, *"Distributed Strategies for Mining Outliers in Large Data Sets"*, IEEE transactions on knowledge and data engineering, Vol. 25, No. 7, July 2013.

[8] Shu Wu and Shengrui Wang, *"Information-Theoretic Outlier Detection for Large-Scale Categorical Data"*, IEEE transactions on knowledge and data engineering, Vol. 25, No. 3, March 2013.

[9] Charu C. Aggarwal, Philip S. Yu, *"Outlier Detection for High Dimensional Data*", in Proceedings of the 2001 ACM SIGMOD international conference on Management of data, 2001.

[10] Charu C. Aggarwal, Philip S. Yu, *"Outlier Detection with Uncertain Data"*, In SIAM, 2008.

[11] Nam Anh Tran, Gerald Shively, and Paul Preckel, *"A new method for detecting outliers in Data Envelopment Analysis"*, Applied Economics Letters, 2008.

[12] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori, *"Statistical Outlier Detection Using Direct Density Ratio Estimation"*, Knowledge and Information Systems. vol.26, no.2, pp.309-336, 2011.

[13] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons.,3 edition.

[14] Rousseeuw, P. and Leroy, A.: 1996, Robust Regression and Outlier Detection.JohnWiley & Sons., 3 edition.

[15] Grubbs, F. E.: 1969, 'Procedures for detecting outlying observ ations in samples'.Technometri cs 11, 1–21

[16] Laurikkala, J., Juhola, M., and Kentala, E.: 2000, *"Informal Identification of Outliers in Medical Data"*. In: Fifth International Workshop on Intelligent Data Analysisin Medicine and Pharmacology IDAMAP-2000 Berlin, 22 August. Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000.

[17] Datta, P. and Kibler, D.: 1995, *"Learning prototypical concept descriptions"*, In:Proceedings of the 12th International Conference on Machine Learning. pp. 158–166, Morgan Kaufmann. DeCoste, D. and Levine, M. B

[18] Skalak, D. B. and Rissland, E. L.: 1990, 'Inductive Learning in a Mixed Paradigm Setting'. In: Proceedings of the Eighth National Conference on Artificial Intelligence, Boston, MA. pp. 840–847

[19] Aha, D. W. and Bankert, R. B.: 1994, 'Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison'. In: Proceedings of the AAAI-94 Workshop on Case-Based Reasoning

[20] Aggarwal, C. C. and Yu, P. S.: 2001, 'Outlier Detection for High Dimensional Data'.In: Proceedings of the ACM SIGMOD Conference 2001.

[21] Faloutsos, C., Korn, F., Labrinidis A., Kotidis Y., Kaplunovich A., and Perkovic D.:1997, 'Quantifiable Data Mining Using Principal Component Analysis'. Technical Report CS-TR-3754, Institute for Systems Research, University of Maryland, College Park, MD.

[22] Ramaswamy, S., Rastogi, R., and Shim, K.: 2000, 'Efficient Algorithms for Mining Outliers from Large Data Sets'. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Dallas, TX, pp.427–438.

[23] Ester, M., Kriegel, H-P., and Xu, X.: 1996, 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, pp. 226–231. AAAI Press.

[24] Knorr, E. M. and Ng, R. T.: 1998, 'Algorithms for Mining Distance-Based Outliers in Large Datasets '. In: Proceedings of the VLDB Conference.NewYork,USA, pp. 392–403

[25] Byers, S. and Raftery, A. E.: 1998, 'Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes'. Journal of the American Statistical Association 93(442), 577–584

[26] Wettschereck, D.: 1994, 'A study of distance-based machine learning algorithms'. Ph.D. thesis, Department of Computer Science, Oregon State University, Corvallis.

[27] Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D.: 2002, 'A Robust Outlier Detection Scheme for Large Data Sets'. Submitted

[28] Rousseeuw, P. and Leroy, A.: 1996, Robust Regression and Outlier Detection.John Wiley & Sons., 3 edition.

[29] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons.,3edition.

[30] Torr, P. H. S. and Murray, D. W.: 1993, 'Outlier Detection and Motion Segmentation'. In: Proceedings of SPIE

[31] UCI Machine Repository - http://archive.ics.uci.edu/ml/datasets/