outliers in large scale categorical unsupervised data sets for real applications.
5) We have combined entropy and dual correlation with attribute weighting resulting into weighted surfeit entropy where entropy computes uncertainty and dual correlation measures mutual information or attribute relation.
6) Different user defined intrinsic and decision parameters are required for outlier detection problem. Thus results are based on correct values of these parameters.

3. Literature Survey

Methods designed for unsupervised outlier detection in categorical data can be grouped into four categories as follows.

A. Proximity Based Methods

In order to understand this concept, it is the method which measures compactness of objects in terms of distance / density. ORCA [10] and CNB [11] are different algorithms for outlier detection in categorical data. ORCA uses hamming distance and CNB uses common neighbor set. These two methods are not useful for high dimensional data because of difficulty in choosing the distance or density as well as high time and space complexity.

B. Ruled Based Methods

Rule-based methods use the concept of frequent items from association-rule mining. This method considers the frequent or infrequent items as a data set. Objects with few frequent items or many infrequent items are more likely to be considered as anomalous objects than others.

Frequent pattern outlier factor [12] and Otey's algorithm [13] are two well known ruled based techniques. The FIB algorithm includes an initial computation of the set of frequent patterns, using a predefined minimum support rate. All support rates of associated frequent patterns are summed up for each object as the outlier factor of this object. While Otey's algorithm, begins with computation of infrequent items from data set. Outlier factor is calculated using the same. Objects with largest scores are treated as outliers.

C. Other Methods

Random walk, Hyper-graph theory [8] methods are implemented using several approaches. In random walk method [14], outliers are the object who has the low probability to combine with neighbor. That means object remain in their state. In method [15] relationships are considered and mutual dependence based local outlier factor is proposed to detect outliers. There are many other methods cluster based local outlier detection method, classification based method.

In literature several methods have been proposed for outlier detection using information theoretic measures.

1) An information theoretic method is proposed for anomaly detection in audit data sets in [2] using measures like entropy, conditional entropy, relative entropy & information gain. This method aims to identify outliers in the univariate audit data set, where regularity is characterized but not the attribute relation.
2) Information theoretic outlier detection in large scale categorical data holoentropy –sum of entropy and total correlation. This method gives optimal solution to outlier detection by using ITB-SP algorithm.

ITB-SP Method computes holoentropy as follows:

$$HLx(y) = Hx(y) + Cx(y) \quad (1)$$

After computing holoentropy, weighted holoentropy and differential entropy is calculated using (2) and (3) to get anomaly set AS from data set.

$$Wx(y) = \sum_{i=1}^{m} Wx(yi)Hx(yi) \quad (2)$$

$$Hx(xo) = Wx(y) - Wx\{xo\}(y) \quad (3)$$

Table 1: Comparison of systems

Parameter	*CNB*	*ORCA*	*FPOF*	*ITB-SP*
Approach	Proximity based	Proximity based	Ruled based	Information theoretic based
Method	Distance	Distance	Item set frequency	High Dimensional categorical data
Input Data Set	Low dimensional categorical data	High dimensional data in random	Low dimensional Numeric data	High dimensional categorical data
Required parameters	M, sim, k	K, M	Minfreq, maxlen, M	Number of outliers o
Output Data Set	outliers	O-outliers	Value of FPOF, FP-outliers	OS-outlier set
Complexity	$O(n^2(k+S(\theta) + q) + n(k + M))$	$O(n^2q)$	$O(n(2^{T-f}))$	$=O(nm)$

Table 1, shows the comparison between different outlier detection methods using parameters like approach, type, input data set, output data set, complexity and user defined parameters.

D. Necessity of modified outlier detection:

Most of the existing systems are depends on user defined parameters and very few methods are dealing with unsupervised categorical data. Therefore a method should exist which will be able to deal large scale categorical data without requirement of any user defined parameter. There is requirement of method which will perform outlier detection using joint correlation between attributes.

4. Problem Formulation

In this section we first look at how entropy and dual total correlation can be used to capture similarity between outlier candidates. We are proposing Weighted Surfeit Entropy and formulate the outlier detection problem.

A. Entropy: Entropy is measure of information and uncertainty of a random variable.

Let X be the set of n objects $\{x1, x2, x3, ,, ... n x\}$, each xi for $1 \le I \le n$ being a vector of categorical attributes $[y_1, Y_2, y_3, ... y_m]^T$ where m number of its attributes. Now based on chain rule of entropy [4], Entropy of y denoted as $Hx(y)$ can be written as follows.

$$Hx(y) = Hx(y1, y2, \ldots, ym)$$

$$= \sum_{i=1}^{m} Hx(yi|yi-1, \ldots .y1)$$
$$= Hx(y1) + Hx(y2|y1) + \ldots + Hx(ym|ym-1, \ldots, y1) \quad (4)$$

Where

$$Hx(ym|ym-1, \ldots, y1) = -\sum_{ym, ym-1, \ldots, y1} p(ym, ym-1, \ldots y1) \log p(ym|ym-1, \ldots y1).$$

Entropy of dataset decreases significantly with removal good outlier candidates.

B. Total Correlation:

It is defined as summation of mutual information of multivariate discrete random vector y, [8,1] and it is denoted as $C_x(y)$. Total correlation is based on Watanabe's proof. Total correlation can be expressed as :

$$Cx(y) = \sum_{i=1}^{m} Hx(yi) - Hx(y) \quad (5)$$

C. Dual total correlation:

The dual total correlation [17] calculates the amount of entropy present in Y beyond the sum of the entropies for each variable conditioned upon all other variables. The dual total correlation is also called as the surfeit entropy and the binding information. In this paper we describe dual total correlation as Ex(Y) and expressed as

$$Ex(y) \equiv \left(\sum_{Xi \in X} Hx \backslash xi(y)\right) - (n-1)Hx(y) \quad (6)$$

Where n is number of attributes.

To weight the entropy of each attribute, we are using a reverse function of the entropy, as follows:

$$Wx(yi) = 2\left(1 - \frac{1}{1+\exp(-Hx(yi))}\right) \quad (7)$$

The weighted Surfeit entropy is defined as follows:

Definition 1: The weighted surfeit entropy EW_X (Y) is the sum of weighted entropy on each attribute of the random vector Y.

$$EWx(Y) = \sum_{i=1}^{m} Wx(yi)Hx(yi) \quad (8)$$

Outliers are detected by minimizing the surfeit entropy through the removal of outlier candidates; Proposed strategy Have weighting the entropy of each individual attribute in order to give more importance to those attributes with small entropy values.

D. Formal definition of outlier detection:

We are using weighted surfeit entropy for outlier detection outliers .We consider that set of outlier candidates is the best if entropy of dataset significantly decreases with its removal from dataset.

Definition 2: X be a given dataset with n objects and a subset Out(o) is defined as the set of outliers if it minimizes the weighted surfeit entropy of dataset X with o objects removed.

E. Differential surfeit entropy:

Definition 3: Given an object x_o of X, the difference of weighted surfeit entropy $e_x(xo)$ between the data set X and the data set X\{xo} is defined as the differential surfeit entropy of the object x_o.

$$ex(xo) = Wx(y) - Wx\{xo\}(y) \quad (9)$$

F. Outlier factor:

Outlier factor is a measure of how likely x_o is an outlier. An object x_o with a large outlier factor value is more likely to be an outlier than an object with a small value. Outlier factor of an object x_o is denoted as $OF(x_o)$ is defined as :

$$OF(xo) = \sum_{i=1}^{m} OF(xo, i) \quad (10)$$

5. Proposed Approach

Our proposed approach is based on weighted entropy and differential entropy which can be calculated using equation (8) and (9). System will take data set file of format .CSV and gives output file with outliers removed.

System Architecture

To address the problem discussed above in need of effective outlier detection in unsupervised data set. A proposed methodology with surfeit entropy and to deal with large scale categorical data is considered as shown in Fig. 1

1) GUI Handler:

It provides following functionality:

- File selector (CSV File)
- Display for Attributes
- Display for Outliers (Outcome)

2) File Processor:

It will handle following tasks:

- Separate objects and attributes.
- Saving outlier results.l

3) Outlier Detector:

It will handle following tasks:

- Calculate Entropy
- Calculate Dual total Correlation
- Calculate weighted surfeit entropy
- Calculate Outlier factor
- Getting outlier set
- Getting data set file with removal of attributes

4) Report generator:
- Generate Report
- Generate comparison model using graphs

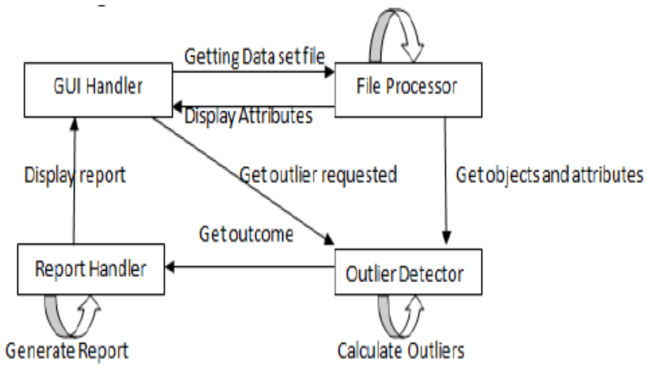

Figure 1: System Architecture

Mathematical model:

The proposed concept is constructed on the assumption that elimination of outliers will improve the purity of data set and reduces EWx(y). When a normal object is removed from the data set, the value of EWx(y) should increase. Thus, the objects with positive $\mathbf{e(x_i)}$ are defined as the outlier candidate set (OS). The objects with non-positive $\mathbf{e(x_i)}$ are defined as elements of normal object set (NS).

$$OS = \{Xi|e(Xi) > 0|\} \; And \; NS = \{Xi|e(Xi) \leq 0|\}$$

SEB-SP Algorithm for outlier detection:

In this section, we have derived Surfeit entropy based single pass greedy algorithm for outlier detection. In this algorithm outlier factors are computed only once, and the o objects with largest values are identified as outliers. This algorithm is parameter-less as we do not need to provide any user defined parameters.

Algorithm: SEB-Single Pass
1. **Input:** data set X
2. **Output:** Outlier set S
3. Compute w(y_0) for (1≤ i ≤ m) by (7)
4. Set OS=**null**
5. **for** i=1 to n do
6. Compute OF(xi) and obtain OS by (10)
7. **end for**
8. Build S by searching in OS

Greedy approach is used to find out outliers from input data set in above algorithm. Algorithm firstly computes weighted entropy for each attribute. Then entropy of each attribute is updated. The attribute entropy is always changes when outliers are detected and removed from the data set, then calculates outlier factor for each attribute and get the largest OF set which will convert to OS (Outlier set). After that set S will be built. Complexity of the algorithm is O (nm), as we are not using any searching algorithm.

6. Conclusion

This paper discusses many outlier detection methods based on information theory. We are proposing novel method which will overcome limitations of previous methods. This paper formulates outlier detection as an optimization problem and proposed a practical, unsupervised, parameter less algorithm for detecting outliers in large-scale categorical data sets. Effectiveness of our approach results from a new concept of surfeit entropy. The efficiency of our algorithms results from the outlier factor function derived from the differential entropy. In particular, in our proposed approach dual total correlation and surfeit entropy works more effectively to remove outlier from large scale categorical attributes.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.
[2] V.J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev., vol. 22, no. 2, pp. 85- 126, 2004.
[3] E.M. Knorr and R.T. Ng, “Algorithms for Mining Distance-Based Outliers in Large Data Sets,” Proc. 24rd Int’l Conf. Very Large Data Bases (VLDB ’98), 1998.
[4] S.R. Gaddam, V.V. Phoha, and K.S. Balagani, “K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 345-354, Mar. 2007.
[5] V. Chandola, Banerjee, and Kumar, “Anomaly Detection: A Survey ”, ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.
[6] V. Hodge and J. Austin “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev, vol. 22, no. 2, pp. 85-126, 2012.
[7] J. Zhang,” Advancements of Outlier Detection: A Survey,” ICST Transactions on Scalable Information Systems, Vol.3, no. 1, March 2013.
[8] Sheu Wu, Member IEEE, and Shengrui Wang, Member IEEE, “Information-Theoretic Outlier Detection for Large-Scale Categorical Data,” IEEE transactions on knowledge and data engineering, vol. 15, no. 3, march 2013.
[9] T. Cover and J. Thomas, “Elements of Information Theory” John Wiley & Sons, 1991, pp.12-21.
[10] Ayman Taha, Ali Hadi,” A General Approach for Automating Outliers Identification in Categorical Data, IEEE conference on Computer System & applications, pp.1 – 8, May 2013.
[11] S. Li, R. Lee, “Mining Distance-Based Outliers from Categorical Data,” Proc. IEEE Seventh International Conference, Data Mining Workshops (ICDM ’07), 2007.
[12] S.D. Bay “Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule,” Proc. Ninth ACM SIGKDD International Conf. Knowledge Discovery and Data Mining (KDD ’03), 2008.
[13] Z. He, X. Xu, Z.J. Huang, “FP-Outlier: Frequent Pattern Based Outlier Detection,” Computer Science and Information Systems, vol. 2, pp. 103-118, 2005.
[14] M.E. Otey, A. Ghoting, “Fast Distributed Outlier Detection in Mixed-Attribute Data Sets,” Data Mining and Knowledge Discovery, vol. 12, pp. 203-228, 2006.

[15]] H.D.K. Moonesignhe, Tan, “Outlier Detection Using Random Walks,” Proc. IEEE 18th Int’l Conf. Tools with Artificial Intelligence (ICTAI ’09), 2009.
[16] W. Qian, H. Lu, and A. Zhou, “Finding Centric Local Outliers in Categorical/Numerical Spaces, Knowledge and Information Systems, vol. 8, no. 3, pp 309-338, 2006.
[17] Nicholas Timme,_ Wesley Alford, Benjamin Flecker, and John M. Beggs,” Multivariate information measures: an experimentalist’s perspective”, 28 Nov 2011.
[18] S. Watanabe, “Information Theoretical Analysis of Multivariate Correlation,” IBM J. Research and Development, vol. 4, pp. 66-82, 1960.
[19] http://www.data setgenerator.com/, 2011.