CATEGORICAL DATABASE INFORMATION-THEORETIC APPROACH OF OUTLIER DETECTION MODEL

Chiranji Lal Chowdhary [1], Abhishek Ranjan [2], D. S. Jat [3]

[1] School of Information Technology and Engineering, VIT University, Vellore, India
[2] Maseru Campus, Botho University, Maseru, Lesotho
[3] Namibia University of Science and Technology, Windhoek, Namibia

Corresponding author: Chiranji Lal Chowdhary, c.l.chowdhary@gmail.com

ABSTRACT: Outlier detection system discovers the novel or rare events, anomalies, vicious actions, exceptional phenomena. It is mandatory to find these anomalies in data mining because the presence of these objects usually makes the database inefficient. An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Finding objects that do not conform to well-defined notions of expected behaviour in a dataset is called outlier detection. Outlier detection is a pre-processing step for locating these non-conforming objects in data sets. This outlier detection is a challenging process in large scale database since it has high dimensional data with low anomalous rate. Here outliers are defined formally and the optimized ways to detect outliers is also proposed here. Optimization in outlier detection is achieved by a new concept of holoentropy which combines entropy and total correlation. It is a more effective and efficient practical phenomenon in outlier detection methods. It can be used effectively to deal with both large and high-dimensional datasets.

KEYWORDS: Outlier Detection, Anomalies, Optimization, Holoentropy, Data Mining.

1. INTRODUCTION

The problem of finding objects in a data set that does not conform to well-defined notions of expected behaviour. The objects detected are called outliers, also referred to as anomalies, surprises, aberrant, etc. The supervised or the semi-supervised approaches all need to be trained before use. In 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. It is difficult to obtain training data sets which covers all possible abnormal behaviour that occurs in a data and this task is highly time and space consuming.

The objects detection is called as Outlier Detection, set of data is known as Object. Large-scale data set, those objects that conform to well-defined notions of expected behavior. Our aim is to detect outliers in unsupervised data set using different outlier detection approaches and to make it Large-scale data set. An effective and efficient methods that can be used to solve the outlier detection problem in real applications is been proposed.

When faced with a large data set with millions of high-dimensional objects and a low anomalous data rate, picking the abnormal and normal objects to compose a good training data set is time-consuming and labour-intensive. The unsupervised approach is more widely used than the other approaches because it does not need labelled information. Non-conforming objects in data sets are the outliers. Experimental results have shown that detecting these outliers from categorical data sets and removing them will make the data sets effective and efficient.

2. RELATED WORK

In data mining discovering novel or rare events, anomalies, vicious actions, exceptional phenomena are mandatory. Objects that do not conform to well-defined notions of expected behaviour in a dataset is called outlier detection. Outlier detection is a pre-processing step for locating these non-conforming objects in categorical data sets. It is a challenging criterion for defining a meaningful similarity measure for categorical data. A formal definition of outliers and an optimization model of outlier detection is been proposed here, via a new concept of holoentropy which combines entropy and total correlation. ITB-SS and ITB-SP are more effective and efficient practical parameter outlier detection methods, than mainstream methods and can be used to deal with both large and high-dimensional data sets [WW09].

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection [CA16]. Formal definition of outliers and an optimization model of outlier detection, using a new concept of holoentropy that takes both entropy

and total correlation into consideration. The current work focuses on finding single records that are anomalous. Sometimes in real world applications we are more interested in detecting groups of unusual records that deviate from the norm, rather than detecting the records separately. We need to specify a similarity measure, and group records on the basis of it. If the data has temporal and/or spatial components, they provide a natural measure for grouping [DS07].

In most applications, anomalies are defined as data points that are 'abnormal'. Quite often we have access to data which consists mostly of normal records, along with a small percentage of unlabelled anomalous records. We are interested in the problem of unsupervised anomaly detection, where we use the unlabelled data for training, and detect records that do not follow the definition of normality. ITB-SS and ITB-SP, ITB-SS – Information Theory-Based Step by-Step ITB-SP - Information-Theory-Based Single-Pass. An extensive survey of anomaly detection techniques developed in machine learning and statistical domains A broad review of anomaly detection techniques for numeric as well as symbolic data is presented. An extensive review of novelty detection techniques using neural networks and statistical approaches has been presented [CBK09].

Mining outliers in database is to find exceptional objects that deviate from the rest of the data set. Besides classical outlier analysis algorithms, recent studies have focused on mining local outliers, i.e., the outliers that have density distribution significantly from their neighbourhood. The estimation of density distribution at the location of an object has so far been based on the density distribution of its k-nearest neighbours. Aim is to propose effective and efficient methods that can be used to solve the outlier detection problem in real applications. HITS is a well-known algorithm developed for analysing hyperlink structures in Web environment. ROCK is a well-known clustering algorithm for datasets with categorical attributes. It takes advantages of common neighbours, based on categorical data similarities, to define links between pairs of objects [Y+06].

Much of the intrusion detection research focuses on signature (misuse) detection; here models are built to recognize known attacks. However, signature detection, by its nature, cannot detect novel at-tacks. Anomaly detection focuses on modelling the normal behaviour and identifying significant deviations, which could be novel attacks. In this paper we explore two machine learning methods that can construct anomaly detection models from past behaviour. The unsupervised approach is more widely used than the other approaches because it does not need labelled information. The local outlier assigns each data a local outlier factor LOF [J+06].

Outlier detection methods for categorical data can be characterized by the way outlier candidates are measured w.r.t. other objects in the data set. In general, outlier candidates can be assessed based either on data distribution or on attribute correlation, which provides a more global measure. They can also be assessed using a between-object similarity or local density, which provides a local measure. To solve the optimization problem, we derive a new outlier factor function from the weighted holoentropy and show that computation/updating of the outlier factor can be performed without the need to estimate the joint probability distribution [Die98].

Discovery of objects with exceptional behaviour is an important challenge from a knowledge discovery standpoint and has attracted much attention recently. In this paper, we present a stochastic graph-based algorithm, called Out Rank, for detecting outlying objects. In our method, a matrix is constructed using the similarity between objects and used as the adjacency matrix of the graph representation. The heart of this approach is the Markov model that is built upon this graph, which assigns an outlier score to each object. We combine entropy and total correlation with attribute weighting to define the concept of weighted holoentropy [FS10].

Here evaluations on a small real data set and a bundle of synthetic data sets shows that the proposed algorithms do tend to optimize the selection of candidates as outliers. Moreover, our experiments on real and synthetic data sets in comparison with other algorithms confirm the effectiveness and efficiency of the proposed algorithms in practice. The unsupervised anomaly detection approach detects anomalies in an unlabelled data set under the assumption that the majority of the objects in the data set are normal. This algorithm assumes that the data is in random order. If the data is not in random order and is sorted then the performance can be poor. Algorithm depends on the independence of examples. It addresses the scaling problem with an algorithm based on randomization and pruning which finds outliers on many real data sets in near linear time [BS03, Cho11].

3. ARCHITECTURE DIAGRAM OF PROPOSED SYSTEM

3.1 Overview of System

Outlier detection system (Fig. 1) is used to remove outliers or aberrant or anomalies from the datasets stored in database. Usually large scale datasets are high dimensional and has low anomalous rate. Since it has millions of data it is not very easy to label any data as outliers or normal data. So unsupervised database is used which does not need any classification as outlier or normal datasets. Outliers are those objects that do not conform to the boundaries or conditions set for those

datasets. These outliers if present will make the database inefficient or ineffective. In order to make the datasets efficient and effective a pre-processing is done to the datasets which finds the outliers. These outliers can be removed from the database. In this system the pre-processing of data is done by analysing the attributes (entropy), shared and dependant information (total correlation) and finally by holoentropy. Entropy and total correlation analyses the entire attribute and shared attribute and categorizes datasets as normal or outliers whereas holoentropy highlights the exact field which causes the whole record to be outlier so that it can be easily identified and removed from the database.

3.2 Outlier Detection on Shared Information

In this module after successful login the User/Admin can view all the personal/professional information stored in database. The datasets which has to be analysed by checking with shared dependent attributes is sent for analysis by choosing Total Correlation. The dependent attributes which has shared information are analysed thoroughly and it is categorized to normal and outlier datasets. The result of analysis for each tuple is produced that can be viewed by the user/admin and the outlier values can be updated with correct values.

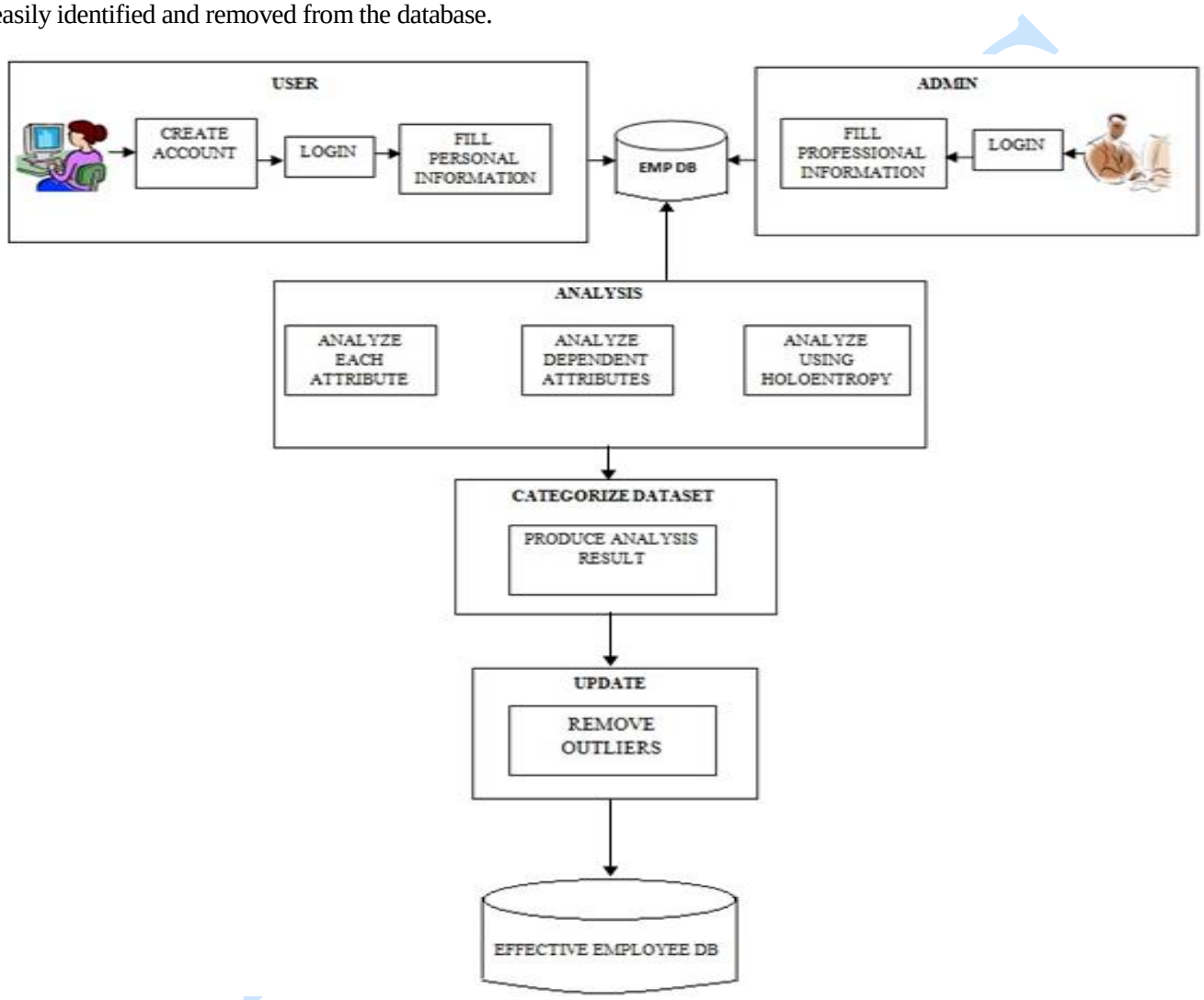

Fig. 1. The Outlier detection system

3.3 Information Theoretic Based Outlier Detetion

In this module User/Admin logs in by giving their username and password. On successful login the User/Admin can view all the personal/professional information stored in database. The field which causes the whole tuple as outlier can be found by choosing "apply holoentropy". The attributes are analysed in every step and pass. The result of analysis is produced which highlights the outlier fields which helps to easily identify the outliers and remove them to make the employee database effective.

4. EXERIMENTAL EVALUATION

4.1 Registration

The user is made to register first. Registration is done by giving the Username and Password. If any of the details are left unfilled then registration process fails. After giving in the user name, Password and Confirming password the user has to click on Register now. When the user gets registered successfully Registered successfully message is displayed. If the user lefts any of the field empty then

Enter all fields message is displayed. The User is restricted to register again with same username. After successful registration the user can sign in to store their personal details.

There exists two login-User login and Admin login. The user can sign in by using the same username and password that they used for registering. The user has to click on Login to store their personal details. If the user gives in proper details for login then login successful message gets displayed and user gets logged in. If the user gives in incorrect credentials to login then Username/Password invalid message gets displayed and login fails. The user can fill in all personal information to store in database. If user wants to register then the sign up button can be used.

The admin can sign in by using admin username and password. If the admin signs in with proper credentials then Admin login successful message gets displayed. If Login is done with incorrect credentials then Login failed message gets displayed. The admin maintains all professional details of employee. The admin fills in the salary and other professional details of employee to store in database. The Exit button can be used if the login process has to be terminated.

After successful login the user can fill in all personal information. The Personal information is stored in database by clicking on Store Data button. On storing the data successfully Stored successfully message gets displayed. To send data for analysis of outliers Examine Personal Information button can be chosen.

Admin fills in all these information of each employee and stores it in database by clicking on Store Salary Details. The Apply Total Correlation button can be used for applying total correlation on the records stored in database. The Apply Entropy button can be used for applying entropy on the records stored in database.

4.2 Attribute Based Outlier Detection

Attribute based outlier detection refers to detecting outliers using entropy. The entropy can be used as a global measure in outlier detection. To detect outlier using entropy each field is analysed separately to find non-matching attribute values.

4.3 User Entropy

The Personal Information can be analysed for its outlier by clicking on Examine Personal Information. All the personal information of the employee are retrieved and displayed on clicking Click to view personal information. The Fields to which entropy is to be applied gets listed (employee name, gender, age, dob, address). The field to which entropy has to apply is selected and Apply Entropy button is clicked. Each field can be selected separately and can be analysed for normal and outlier dataset. The dataset of the selected field is sent for analysis. After the dataset is sent for analysis successfully Data Sent message gets displayed (Fig. 2).

Fig. 2. User Personal Information

Each dataset that is received is analysed if it is of null value and also checked for its type and categorized into normal data and outlier data accordingly. The analysis of dataset is done by clicking on Click Here button. After analysing the dataset completely Analysis Successful message gets displayed. If the dataset is of proper datatype it is put under normal dataset if not the dataset is tagged as outlier (Fig. 3).

Fig. 3. User Entropy-Gender Result

The address attribute given in personal information is analysed for its outliers and it is categorized to normal and outliers. The Employee name attribute is analysed for outliers or anomalies by applying entropy conditions to it. If the employee name is left empty or if the name has invalid character it is then tagged as outlier. Else it is given as normal dataset (Fig. 4).

Fig. 4. User Entropy-Employee Name Result

4.4 Admin Entropy

The professional information can be analysed for its outlier by clicking on Apply Entropy button. All the professional information of the employee are retrieved and displayed on clicking View Information button. The Fields to which entropy is to be applied gets listed (Location, position, qualification, department name). The field to which entropy has to apply is selected and Apply Entropy button is clicked. Each field can be selected separately and can be analysed for normal and outlier dataset. The dataset of the selected field is sent for analysis. The dataset is sent for analysis (Fig. 5).

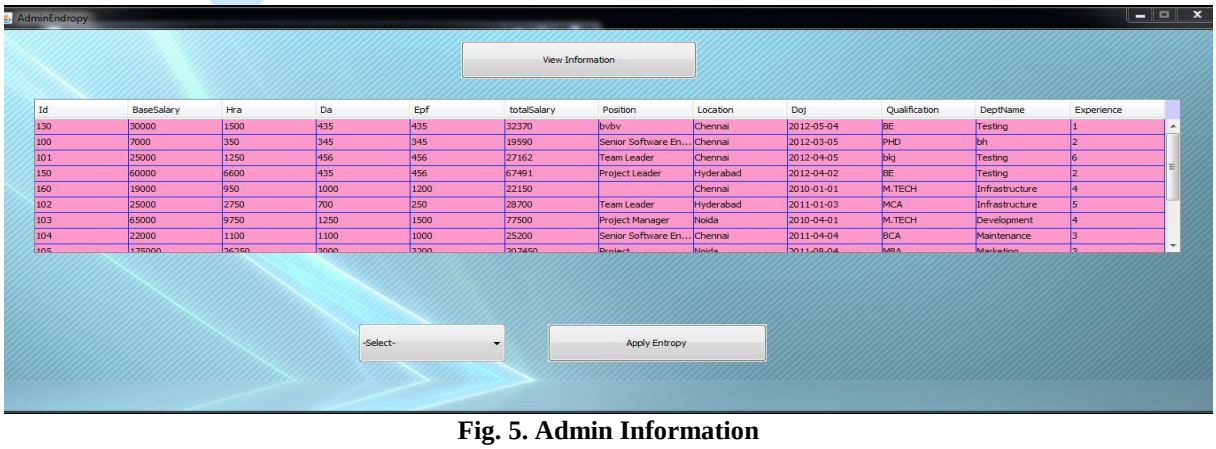

Fig. 5. Admin Information

Each dataset that is received is analysed if it is of null value or if it has values other than listed values and also checked for its type and it is categorized into normal data and outlier data accordingly. The analysis of dataset is done by clicking on Click Here button. After analysing the dataset completely Analysis Successful message gets displayed. If the dataset is of proper datatype it is put under normal dataset if not the dataset is tagged as outlier (Fig. 6).

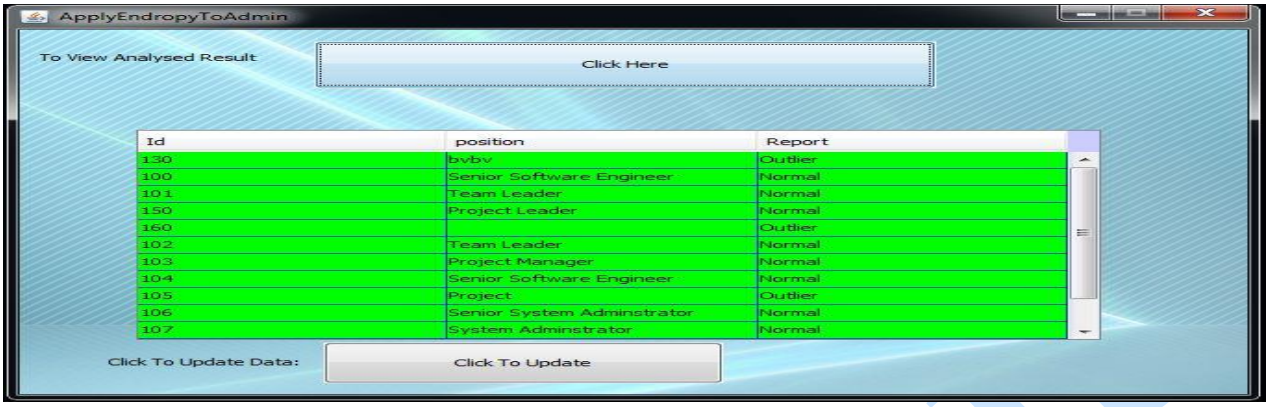

Fig. 6. Admin Entropy-Position Result

The Location attribute is sent for analysis to find outliers in dataset by applying entropy conditions. If the location is left empty or if the location has unmatched character types or if it has values other than a few values it is tagged as outlier. Else it is tagged as Normal dataset. The categorized result of normal and outlier data is produced as result set (Fig. 7).

Fig. 7. Admin Entropy-Location Result

4.5 Outlier Detection On Shared Information

Analysis for outliers is done also on shared information or mutually dependent information. The total correlation is a quantity that measures the mutual dependence or shared information of a data set.

4.6 User Total Correlation

The personal information stored in database used to remain dependent with each other. So Total correlation is applied for each record to find if there exist any outliers in dependent or shared information. Apply Total Correlation button is clicked to initiate the analysis of dataset for outliers. The data is sent to next step for analysing outliers. After sending the data successfully data sent message gets displayed. The Examination of outliers on each record is done by clicking on Click Here button. After successful analysis, Analysis Successful message gets displayed and the dataset gets categorized to normal and outlier dataset based on different conditions.

4.7 Admin Total Correlation

The professional information stored in database by admin used to remain dependent with each other. So Total correlation is applied for each record to find if there exist any outliers in dependent or shared information. Apply Total Correlation button is clicked to initiate the analysis of dataset for outliers. The Examination of outliers on each record is done by clicking on Verify button. After clicking on it the records gets categorized to normal and outlier dataset based on different (Fig. 8).

Fig. 8. Admin Total Correlation

4.8 Information Theoretic Based Outlier Detection

Holoentropy effective and efficient method to determine the outliers present in the large scale dataset. It combines the result of Entropy and Total Correlation to detect the outliers or anomaly present in the datasets. The above two methods, namely Entropy and total correlation check the attributes or dependent attributes and categorizes data to normal and outliers whereas holoentropy check all attributes in every single step and pass and highlights the field which causes the whole record to be outlier.

4.9 User Holoentropy

The personal information is analysed to find if there exist any outliers. Apply Holoentropy button is clicked to initiate the analysis of dataset for outliers. The Result of analysis is produced on clicking Holoentropy Results button. The result set produced highlights the outlier field which helps in easier identification of outliers. To remove the outliers found and update with normal values Update button can be used (Fig. 9).

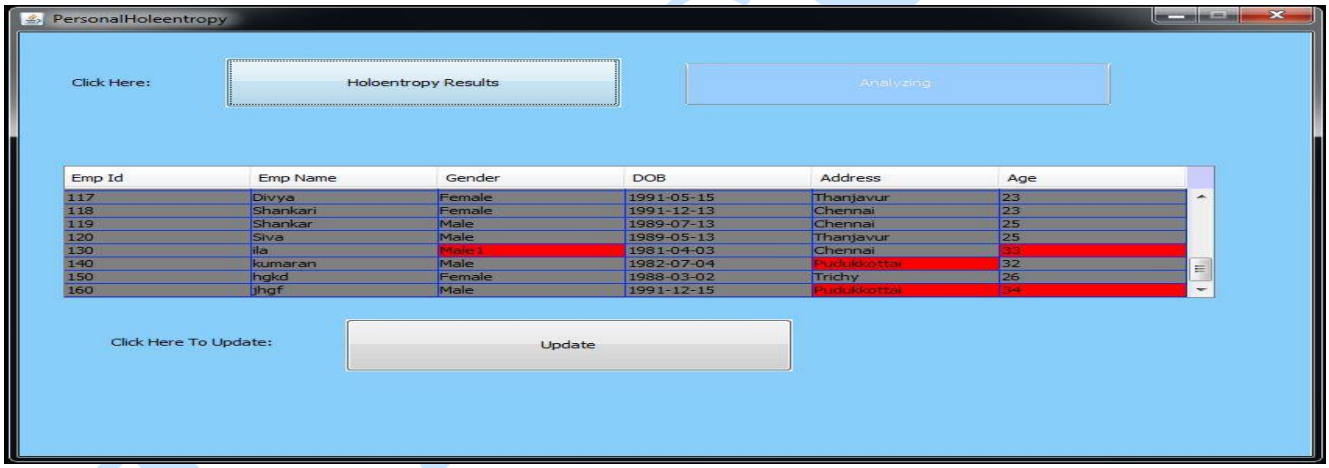

Fig. 9. User Holoentropy

4.10. Admin Holoentropy

The professional information is analysed to find if there exist any outliers. Apply Holoentropy button is clicked to initiate the analysis of dataset for outliers. The Result of analysis is produced on clicking Holoentropy Results button. The result set produced highlights the outlier field which helps in easier identification of outliers. To remove the outliers found and update with normal values Update button can be used (Fig. 10).

5. CONCLUSIONS

This work mainly focuses on detecting outliers, rare events, anomalies, vicious actions, exceptional phenomena that occurs in database. A large scale database usually has millions of high-dimensional objects. These objects can have low anomalous data rate. Usually picking the abnormal and normal objects to compose a good training data set is time-consuming and labour-intensive. So, large scale databases usually use unsupervised approach because it does not need labelled information. These datasets can also have outliers in it. The outliers can be detected by examining attributes using entropy, shared and dependent attributes using total correlation. The detection of outliers is made more efficient by using holoentropy. Detecting these outliers from categorical data sets and removing them will make the data sets effective and efficient.

Fig. 10. Admin Holoentropy

REFERENCES

[BS03] **S. D. Bay, M. Schwabacher** - *Mining distance-based outliers in near linear time with randomization and a simple pruning Rule*, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

[Cho11] **C. L. Chowdhary** - *Linear feature extraction techniques for object recognition: study of PCA and ICA*, Journal of the Serbian Society for Computational Mechanics, vol. 5(1): 19-26, 2011.

[CA16] **C. L. Chowdhary, D. P. Acharjya** - *A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique*, International Journal of Healthcare Information Systems and Informatics, 11(2): 38-61, 2016.

[CBK09] **V. Chandola, A. Banerjee, V. Kumar** - *Anomaly detection: a survey, ACM Computing Surveys*, vol 41(3): 15:1-15:58, 2009.

[Die98] **T. G. Dietterich** - *Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation*, vol. 10(7): 1895-1923, 1998.

[DS07] **K. Das, J. Schneider** - *Detecting anomalous records in categorical data sets*, in Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.

[FS10] **M. Filippone, G. Sanguinetti** - *Information theoretic novelty detection*, Pattern Recognition, vol. 43(3): 805-814, 2010.

[J+06] **W. Jin, A. K. H. Tung, J. Han, W. Wang** - *Ranking outlier using symmetric neighbourhood relationship*, in Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining, 2006.

[WW09] **S. Wu, S. Wang** - *Information-theoretic outlier detection for large-scale categorical data*, IEEE Transaction on Knowledge and Data Engineering, vol. 25(3): 589-602, 2013.

[Y+06] **J. X. Yu, W. Qian, H. Lu, A. Zhou** - *Finding centric local outliers in categorical / numerical spaces, knowledge and information systems*, vol. 9(3): 309-338, 2006.