ISSN 2047-3338

Null Value Estimation in Web Environment by Using Fuzzy Rule Based K-Mean Clustering

Swati Agrawal
Department of Computer Science, All Sents Engineering College, Bhopal
swati_cse@yahoo.com

***Abstract*– In the Web environment web log file capture operational data generated through internet for analysing user's browsing behaviour and many other security issues. The captured operational data is useful for build use profile, web designing and acts as evidence in web forensic and many other security issues. In real world there are lots systems that participate in web environment having incomplete information because of that web log file affected through noise which lead many of inconvenience. Estimation and handling of these noises in web log is major issue in web forensic and other web related security issue. For evaluating that incomplete information null value estimation is very precious technique. This paper proposed a null value estimation technique based on fuzzy rule based k-means algorithm to deal with that noise. Proposed technique enhances the performance of k-means clustering algorithm by encapsulating advantage of fuzzy rule over that.**

***Index Terms*– Web Mining, Web Log, Null Value and Log Parser**

I. INTRODUCTION

WEB log mining is application of data mining techniques to discover user access patterns from web data. Web usage/log data captures web-browsing behavior of users from a web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is Access logs on server side, which keeps information about user navigation [1]. Handling Null values in web server log, however, is of a more recent origin with somewhat different constraints and objectives.

For improve the performance of database by handling null values. The possible approach is uses the concept of clustering by the trained methodology, which also performs on the real log record [7].

The rapid expansion of the World Wide Web has created an unprecedented opportunity to disseminate and gather information online. As more data are becoming available, there is much need to study web-user behaviors to better serve the users and increase the value of enterprises. One important data source for this study is the web-log data that traces the user's web browsing [2]. There is much need to study handling null values, based on the web-log data. The results of handling null values can be used for strengthen products to the customer, investigate the client activities, suggesting gathering of evidences in web server log and handling noisy data on the web.

II. RELATED WORK

In [1] Ching-Hsue Cheng, Liang-Ying Wei, Tzu-Cheng Lin attempt to Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value. In this paper, they present a new method for estimating null values in relational database systems based on adaptive learning techniques. The proposed method utilizes clustering algorithms to cluster data, and calculate coefficient values between different attributes by generating minimum average error. To verify our method, this paper utilizes two database (Waugh's database and Employee relational database), and Mean of Absolute Error Rate (MAER) as evaluation criterion to compare with the listing methods. Therefore, we propose the degree of influence to estimate null values in relational databases, based on clustering and the coefficient values generating by adaptive learning techniques.

In [4] Muhammad Nazrul Islam, Pintu Chandra Shill, Muhammad Firoz Mridha, Dewan Muhammad Sariful Islam and M.M.A Hashem attempt to Generating Weighted Fuzzy Rules for Estimating Null Values Using an Evolutionary Algorithm. In their paper they present the technique to generate weighted Fuzzy rules from relational database systems for estimating null values using Evolutionary algorithms. The parameters (operators) of the Evolutionary algorithms are adapted via Fuzzy systems. They have fuzzified the attribute values using membership functions shape, type and parameter values. The results of the Evolutionary algorithms are the weights of the attributes. The different weight of attribute generates a set of Fuzzy rules. From this they have obtained a set of rules. Their proposed techniques have a higher average estimated accuracy rate and able to estimate the null values in relational database systems.

In [7] Ching-HsueCheng and Jia-WenWang attempt to analyze database to estimate null values, in this, they proposed a new approach to estimate null values in relational database, which utilize other clustering algorithm to cluster data, and use fuzzy correlation and distance similarity to calculate the correlation of different attribute. In this paper, they present two kinds of correlation method to estimate null values in relational database based on fuzzy correlation and distance similarity. It can estimate null values in relational database systems more accurately.

III. Proposed Work

In the Web environment the captured end user interaction from Web sites and portals may be used to figure out end user behaviour and build user profile, and then perform personalized services. But there is possibility of some incomplete information ie null value in log record due to which it is not possible to investigate the client activities in web environment. In order to overcome these problem presented paper proposed an algorithm with a highly estimated accuracy rate by encapsulating the reward of the fuzzy rule over k-means clustering algorithm for more accurate centre point selection and apply over trained data set and relational database for estimation all together. Whole procedure is to be simulated by using client server architecture having 20 clients and one server and capture web log record.

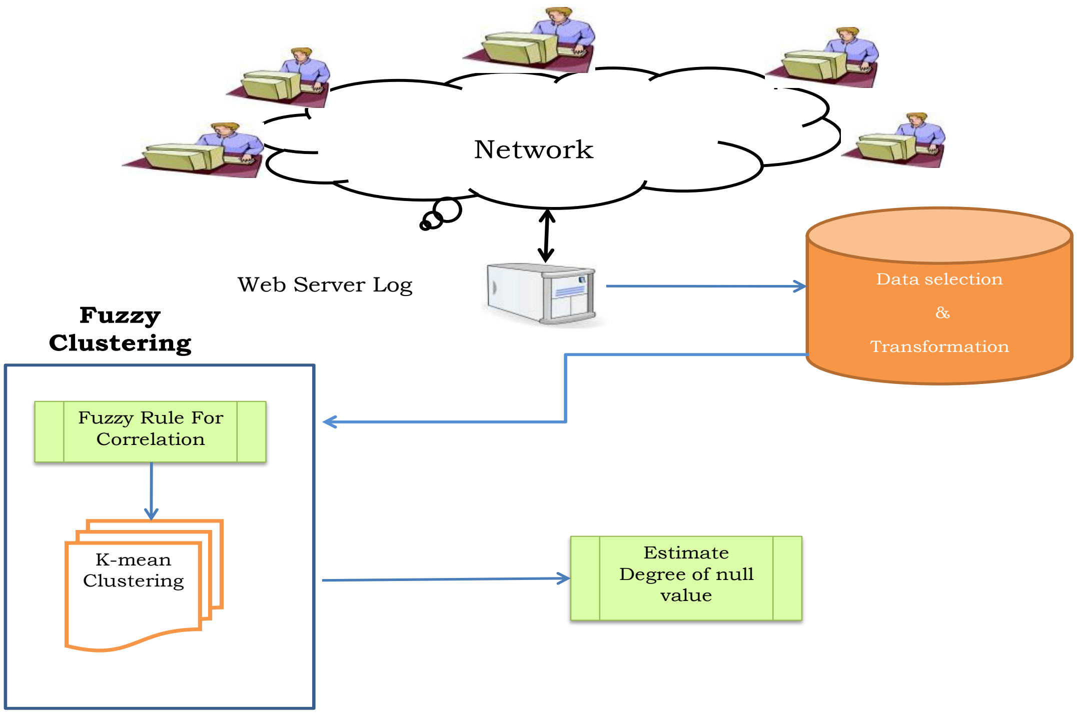

Figure 1: Proposed frame work for estimation of null vale

Major phases of our proposed architecture are Log Records, Parsing of log record, Database of log record, Data Selection and Transformation, Clustering Techniques, Degree of correlation and Estimation of Null Values.

- *Log Records*

Web servers collect information about visits to your website in log files. These are called web server logs. Every time a visitor enters the website an entry is recorded in the log file. Every time a link is clicked an entry is recorded in the log file. The web server log provides you with a history of every click that happens on the site [8].

- *Parsing of log record*

In its raw form, the log file is a big text file. It does not really mean much to look at it. That is where Web Analytics comes in. The software will splice up the log file into discrete pieces of meaningful information and store it in a database. Once the key information is in the database, the software is able to analyze it to identify patterns in the data and generate reports. The process of slicing up a text file into meaningful chunks of information is called parsing [8].

Give this web log data file as an input to web log parsing tool: Web Log Explorer take the log file, parse it and build reports by grouping or filtering the extracted data. Then explore page view request in resulting table and export this report in CSV File (Excel) [10].

The chief advantage of log file parsing, over page tagging, is in its ability to accurately report on website diagnostics. Diagnostic information, such as failed page loads, is found in the web server logs. Page tagging does not have access to it.

❖ *Database of log record*

A database consists of an organized collection of data for one or more multiple uses. One way of classifying databases involves the type of content, for example: bibliographic, full-text, numeric, and image. Other classification methods start from examining database models or database architectures. A database can be understood as a collection of related files. How those files are related depends on the model used. A relational database matches data by using common characteristics found within the data set. The resulting groups of data are organized and are much easier for many people to understand. In computing, databases are sometimes classified according to their organizational approach. The most prevalent approach is the relational database, a tabular database in which data is defined so that it can be reorganized and accessed in a number of different ways[10]. A distributed database is one that can be dispersed or replicated among different points in a network. An object-oriented programming database is one that is congruent with the data defined in object classes and subclasses.

❖ *Data Selection and Transformation*

Log record data base contains the all the information of the log record like the entry point, exit point, visitor stats, referrer stats, user agent state etc. The whole data base contains the many levels information and in a huge amount. It is not possible to perform the same operation on the all category data so we take a particular type of information from this whole database. In our case we take the exit point database. While you can take the different it is depend upon the type of operation which you have to perform. The next step is that the data is feasible to perform the operation. In our case data is available in the text form while we are interested in performing the clustering operation and it is only perform on the numeric data so there is need of transform this textual data into the numeric form in order to perform the clustering [9], [10].

❖ *Clustering Techniques*

Clustering is the processes of organizing data items into groups where members are possess some similarity among them. A cluster is therefore a collection of similar data items or a collection of dissimilar data items belonging to different clusters.

K-means Clustering: The K-means clustering is one of the initials clustering algorithms proposed, it is one of the easiest type of unsupervised learning techniques that solve the clustering problem. It allows the partition of the given data into the k-clusters, the number of clusters is previously decided, after that each cluster randomly guesses centre locations and each data item finds out which centre it is closest to, thus each centre owns a set of data items, now each centre finds its centroid and jumps there, this process is repeated until terminates.

It is one of the initials clustering algorithms proposed.

It allows the partition of the given data into the k-clusters, the number of clusters is previously decided.

1. Each cluster randomly guesses their centre locations.
2. Data item finds out which centre it is closest to.
3. Each centre owns a set of data items.
4. Each centre finds its centroid and jumps there.
5. This process is repeated until terminates.

Although it has the advantage that it is easy to implement, it has two drawbacks. First, it is really slow as in each step the distance between each point to each cluster has to be calculated due to this phenomenon it is expensive in a large dataset. Secondly, this method is very susceptible because we provide the initial value of the clusters. Also It is really slow as in each step the distance between each point to each cluster has to be calculated due to this phenomenon it is expensive in a large dataset [3], [7].

Fuzzy k-Means Clustering: In Fuzzy K-Means a data is formed into K-clusters with every data value in the dataset belongs to all clusters with certain degree. It lies under the unsupervised method and is inherited from fuzzy logic; it is capable of solving the multiclass and ambiguous clustering problems. Fuzziness measures the degree to which an event occurs due to this we are able to increase the probability as respective to the normal probability calculation. In the traditional clustering we assign the each data item to only one cluster. In this clustering we assign different degrees of membership to each point. The membership of a particular data item is shared between various clusters. This creates the concept of fuzzy boundaries which differs from the traditional concept of well-defined boundaries. The well-defined boundary model does not reflect the description of real datasets. This contention led a new field of clustering algorithms based on a fuzzy extension of the least-square error criterion.

In this clustering data is formed into K-clusters with every data value in the dataset belongs to all clusters with certain degree. Fuzziness measures the degree to which an event occurs due to this we are able to increase the probability as respective to the normal probability calculation. The membership of a particular data item is shared between various clusters. This creates the concept of fuzzy boundaries which differs from the traditional concept of well-defined boundaries. The well-defined boundary model does not reflect the description of real datasets.

❖ *Degree of correlation*

Through the coefficient of correlation, we can measure the degree or extent of the correlation between two variables. On the basis of the coefficient of correlation we can also determine whether the correlation is positive or negative and also its degree or extent [8].

Perfect correlation: If two variables changes in the same direction and in the same proportion, the correlation between the two is perfect positive. So the coefficient of correlation in this case is +1. On the other hand if the variables change in the opposite direction and in the same proportion, the correlation is perfect negative. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

Absence of correlation: If two series of two variables exhibit no relations between them or change in variable does not lead to a change in the other variable, then we can firmly say that there is no correlation or absurd correlation between the two variables. In such a case the coefficient of correlation is 0.

Limited degrees of correlation: If two variables are not perfectly correlated or is there a perfect absence of correlation, then we term the correlation as Limited correlation. It may be positive, negative or zero but lies with the limits.

In statistics, correlation and dependence are any of a broad class of statistical relationships between two or more random variables or observed data values. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. Correlations can also suggest possible causal or mechanistic relationships; however, statistical dependence is not sufficient to demonstrate the presence of such a relationship.

❖ *Estimation of Null Values*

Inaccurate, inconsistent or the null data in the database can obstacle research's ability to discover useful knowledge. An effective data quality strategy can help researcher find knowledge in database, allowing them to make right decision and reduce costly operational inefficiencies. In addition, the handling of null values is the major task in data preprocessing of the data mining. But this sort of database system will not operate properly if there are any null values of attributes (incomplete datasets) in the system. In this paper we have been trying to estimate null values from relational database systems. At present some methods exist to estimate null values from relational database systems. The estimated accuracy of the existing methods is not good enough. We use advance technique for estimating null values in relational database systems [4], [7].

IV. Result Analysis

The handling of null values is the major task of the data quality. To overcome time problem proposed methodology estimating null values in web log after parsing in relational database systems based and evaluate using trained data set. The proposed method utilizes fuzzy rule based k means clustering algorithms to cluster data. To verify our method, this paper utilizes same database with the other clustering methods; it is shown that our proposed method is better than the listing methods for estimating null values in relational database systems .Comparison graph between existing technique like k-means Clustering and other with proposed technique is show in Figure 4 where as Figure 2 and Figure 3 show the individual performance of existing K-means clustering technique and modified fuzzy K-means clustering.

As shows in Figure 2 to Figure 4 number of iteration and error rate of proposed fuzzy rule based K-means is gradually decease as compare to existing K-means clustering Algorithm in order to estimate Null value in web log file.

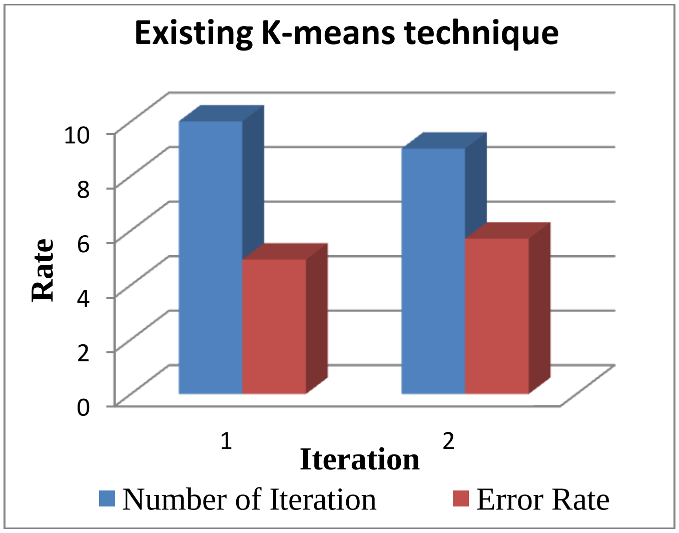

Figure 2: Performance of Existing K-means

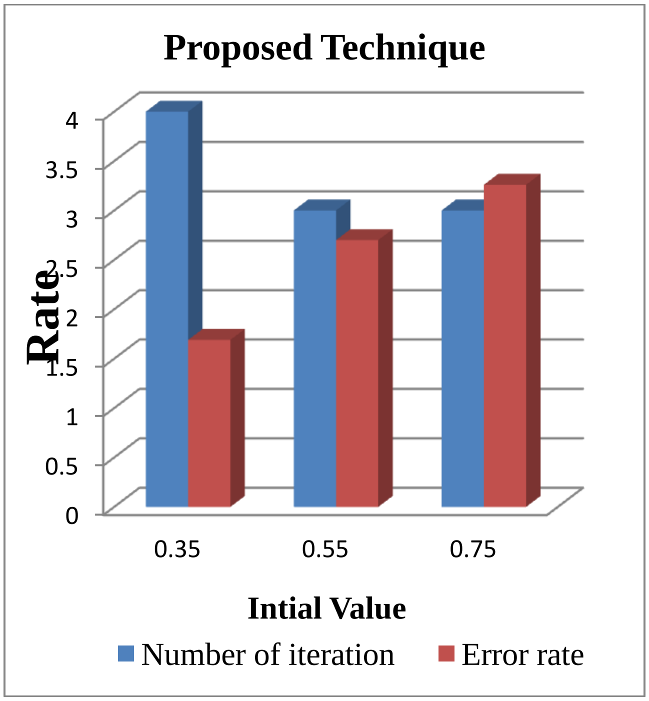

Figure 3: Performance of proposed fuzzy rule based K-means

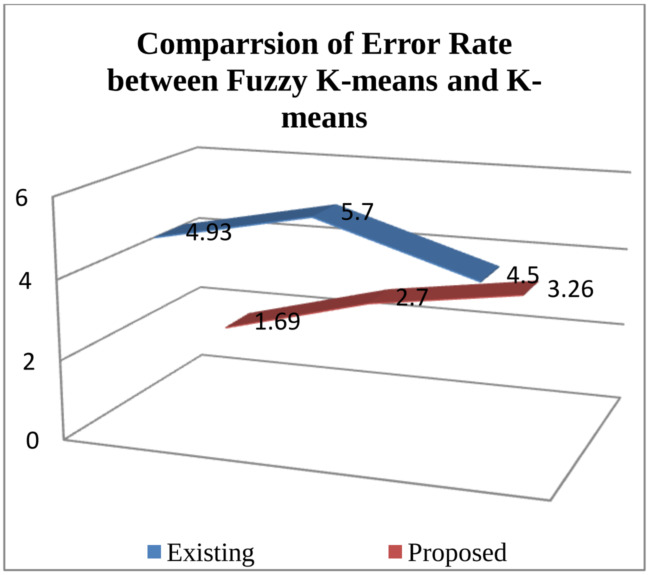

Figure 4: Performance Comparison b/w Existing K-means and fuzzy rule based K-means

Along with that there are some more advantage of proposed technique over simple K-means Clustering ie K-means required number of cluster and seed value manually while proposed technique have no need of giving the seed and cluster value. Here clearly shown that our proposed method have better results than the implemented method.

V. Conclusions

Data mining is one of the popular fields which extracts knowledge from data sets, and plays an important role in computational intelligence, by the log records, investigator can keep an eye on the client's activities but most of the time links break due to which it is not possible store entire log record entry in the database, this creates many problems for investigation of the client activities, to enhance the investigation one of the significant process is estimate the null values. This paper proposed fuzzy rule based k means clustering technique to make a best estimation of null value in web log file and effectively achieve better performance on both relational data base and rough set like log file.

Acknowledgment

I would like to say thanks to my guide who gives their knowledge and time in order to complete this paper. This paper will never complete without the support of faculty members of CSE department.

References

[1] Ching-Hsue Cheng, Liang-Ying Wei, Tzu-Cheng Lin, "Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value", IEEE Innovative Computing, Information and Control, 2007. ICICIC apos; 07. Second International Conference on Volume, Issue, 5-7 Sept. 2007.

[2] S.M. Chen, and H.R. Hsiao, "A new method to estimate null values in relational database systems based on automatic clustering techniques", IEEE Information Sciences: an International Journal, 169, 2005, pp. 47-60.

[3] Shin-Jye Lee, Xiaojun Zeng, "A Modular Method for Estimating Null Values in Relational Database Systems" Eighth International Conference on Intelligent Systems Design and Applications, Nov. 2008.

[4] Muhammad Nazrul Islam, Pintu Chandra Shill, Muhammad Firoz Mridha, Dewan Muhammad, Sariful Islam and M.M.A Hashem, "Generating Weighted Fuzzy Rules for Estimating Null Values Using an Evolutionary Algorithm" in 4th International Conference on Electrical and Computer Engineering, 19-21 Dec 2006.

[5] Claude Rubinson, "Nulls, three-valued logic, and ambiguity in SQL: critiquing date's critique," ACM SIGMOD Record, v.36 n.4, p.13-17, December 2007.

[6] WebLogExpert, http://www.weblogexpert.com accessed on 10/10/201

[7] Imielinski, T. and W. Lipski, "On Representing Incomplete Information in a Relational Database," Proc. 7th Znt. Conf. on Very Large Data Bases, 1981, pp. 388-397.

[8] Nikhil Kumar Singh, Deepak Singh Tomar, Bhola Nath Roy, "*An approach tounderstand the end user behavior through log analysis*" International Journal of Computer Applications (0975 – 8887), August 2010.

[9] Muhammad Kamran Ahmed, Mukhtar Hussain and Asad Raza "*An Automated User Transparent Approach to log Web URLs for Forensic Analysis*" Fifth International Conference on IT Security Incident Management and IT Forensics 2009.

[10] Karen Kent and Murugiah Souppaya, "*Guide to Computer Security Log Management*", Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, 2006