A MAHALANOBIS DISTANCE APPROACH TO RECORD DE-DUPLICATION

[1]V.Rhytthika, [2]K.Manivannan

[1,2] *Computer Science Engineering, PSNA College of Engineering and Technology, (India)*

ABSTRACT

Nowadays, performance of data mining is depends on the data and data mining functions. Various problems occur due to duplication and error in the data and it affect the mining quality. The main objective of this paper is to improve the quality of the database for any kind of DBMS where it increases the performance of data mining operations. Earlier approaches discussed in the literature surveyed to provide solution for anomaly detection and error correction in a DBMS. In this paper data duplication is to find and remove using the TPR-[True Positive Ratio], Error in data is to find and rectify using the FNR – [False Negative Ratio] obtained from Mahalanobis Distance Method.

***Keywords*: Data Duplication, Error Correction, DBMS, Data Mining, TPR, FNR.**

I INTRODUCTION

In database management system the structure of the data, entities available in a same column, data type of the entities should follow the DB rules. It makes perfection on the database tasks like data retrieval, data searching, and data analyze etc. The perfectness of the result of the database functionality becomes poor, due to the data anomalies found in the database.

Detection of anomaly means finding a specific pattern in the data which cannot perform the user expectation and are treated as irrelevant patterns. These kind of irrelevant patterns are called as anomalies, peculiarities and outliers etc. The anomaly detection can be used in advanced applications like credit card fraud, health-care-DB, insurance IDS for cyber security. The most important use of anomaly detection in a database is difficult while apply data actions. Example, in image processing the anomaly detection indicates the tumor presence. In credit card fraud anomalies identify the thief. In past decade a various number of anomalies detection methods have been specially proposed for a specific kind of applications under some domains.

Data duplication is a technique for reducing the amount of storage space an organization needs to save its data. Same file may be saved in several different places by different users or two or more files that are identical may still include much of the same data. De-duplication eliminates these extra copies by saving just one copy of the data. The simplest form de-duplication takes place on the file level that is it eliminates duplicate copies of the same file. This kind of de-duplication is called File-level de-duplication or single instance storage (SIS). It also takes place on the block level, eliminating duplication blocks of data that occur in non-identical files. Block level de-duplication frees up more space than SIS. De-duplication reduces the amount of disk or tape we use and it can reduce storage requirement up to 95 percent. It reduces the amount of network bandwidth required for backup process and some cases it can speed up the back up or recovery process. Storage-based data De-Duplication reduces the amount of storage needed for a given set of files. It is most effective in applications where many copies of very similar or even identical data are stored on a single disk—a surprisingly common scenario. Network data De-Duplication is used to reduce the number of bytes that must be transferred between endpoints, which can reduce the amount of bandwidth required. Virtual servers benefit from De-Duplication because it allows nominally separate system files for each virtual server to be coalesced into a single storage space. At the same time, if a given server customizes a file, De-Duplication will not change the files on the other servers—something that alternatives like hard links or shared disks do not offer. Backing up or making duplicate copies of virtual environments are similarly improved. Our contribution of the work is

Make the data should be error free by pre-processing and normalization. Normalized data is divided into sub windows using Windowing system and finally . Using Mahalanobis distance method the duplicate data is marked and dumped in a separate pool area for further activities [e.g. de-duplicated or eliminated from table].

II RELATED WORK

In [2] the author proposed an approach to data reduction. This data reduction functions are very essential to machine learning and data mining. An agent based population algorithm is used for solving data reduction. Only data reduction is not only the solution for improving the quality of databases. Various size of database is used to provide high classification among the data to find out anomalies. Two algorithms such as evolutionary and non-evolutionary are applied and the results are compared for finding the best suitable algorithm for anomaly detection in [3]. N-ary relations are computed to define the patterns in the dataset [4] where it provides relations in one-dimensional data. DBLEARN and DBDISCOVER [5] are two system developed to analyze RDBMS. The main objective of the data mining technique is to detect and classify data in huge set of database [6] without negotiating the speed of the process. PCA is used for data reduction and SVM is used for data classification in [7]. In [8] the data redundancy method is explored using mathematical representation. Software developed with safe, correct and reliable operations for avionics and automobile based database systems [9]. A statistical QA-[Question Answer] model is applied to develop a prototype to avoid web based data redundancy [10]. GDW-[Geographic Data Warehouses] [11], SOLAP (Spatial On-Line Analytical Processing)is applied for Gist database and other spatial database analysis, indexing, and generating various set of reports without any error.In [12], an effective method was proposed for P-2-P sharing data. During the data sharing the data duplication is removed using the effective method. Web entity data extraction associated with attributes of the data [13]can be obtained using a novel approach which uses duplicated attribute value pairs.

III MAHALANOBIS DISTANCE

The proposed approach utilizes the functionality of Mahalanobis Distance method for finding the difference among entities in each record in a database. This difference indicates the similarity index among two data entities can decide the duplication. In this case, if the distance among two data entities [A and B] is less than a threshold value α, then A and B are decided as duplicate.

A Database DB is a rectangular table T consists of N number of Records R as:

$$\mathbf{R= \{R_1,R_2,.......R_N\}} \qquad \mathbf{(1)}$$

And each record R_i has M number of Entities E as:

$$\mathbf{R_{ij}} = \left\{ \begin{matrix} \mathbf{E_{11}} & \mathbf{E_{12}} & \mathbf{E_{1M}} \\ \mathbf{E_{21}} & \mathbf{E_{22}} & \mathbf{E_{2M}} \\ & & \\ \mathbf{E_{N1}} & \mathbf{E_{N2}} & \mathbf{E_{NM}} \end{matrix} \right\} \quad \mathbf{(2)}$$

E_{ij} is the entity at rth row and jth column in the data. Here i represent the rows and j represents the column. The threshold value α is user defined very small value among 0 and 1. . The data base may be in any form like ORACLE, SQL, and MY-SQL, MS-ACCESS or EXCEL.

Proposed system focuses on applying a Mahalanobis distance metrics indicating similar or dissimilar. MD is a measurement among two data in a database well-defined by appropriate features. Since it accounts for unequal variances as well as correlation among features it will adequately evaluate the distance by assigning different weights or importance factors to the features of data entities. The inconsistency of data can be removed in time series trading information.

Assume two set of groups G_1 and G_2 having data about girls and boys in a school. Let X number girls are categorized as same sub-group in G_1 since their attribute or characteristics are same. It is computed by MD as

$$\mathbf{X = (Xi - Xj) \leq 1} \qquad \mathbf{(3)}$$

The correlation among dataset is computed using Mahalanobis distance.

Data entities are the main objects of data mining. The data entities are arranged in an order according to the attributes. The data set X with M number of attributes is considered as K-dimensional vector and it is represented as:

$$\mathbf{X = (x_{1,} x_{2\ldots\ldots,} x_k)} \qquad \mathbf{(4)}$$

N number data entities $_{Xi}$ form a set

$$\mathbf{D = (X_{1,} X_{2\ldots} X_N) < \beta R^K} \qquad \mathbf{(5)}$$

is known as data set. D can be represented by an N x K matrix

$$\mathbf{D = (x_{ij})} \qquad \mathbf{(6)}$$

Where x_{ij} is the jth component of the data set x_j there are various methods used for data mining. Numerous such methods, for example NN-classification techniques, cluster investigation, and multi-dimensional scaling methods, are based on the processes of similarity between data. As a replacement for measuring similarity, dissimilarity among the entities too will give the same results. For measuring dissimilarity one of the parameters that can be used is distance. This category of measures is also known as reparability, divergence or discrimination measures.

A distance metric is a real-values function d, such that for any data points x, y, and z:

$$\mathbf{d\ (x,y) \geq 0, \text{ and } (x,y) = 0, \text{if and only if } x=y} \qquad \mathbf{(7)}$$

$$\mathbf{d\ (x,y) = d(y,x)} \qquad \mathbf{(8)}$$

$$\mathbf{d\ (x,z) \leq d(x,y) + d(y,z)} \qquad \mathbf{(9)}$$

The first line (9), positive definiteness, assures the distance is a non-negative value. The distance can be zero is for the points to be the same. The second property indicates the symmetry nature of distance. There are various distance formulas are available like Euclidean, manhattans, Lp-Norm and Mahalanobis. In the proposed approach Mahalanobis distance is taken as the main method to find the similarity distance among two data sets.

The Mahalanobis distance is the distance between an observation and the center for each group in m-dimensional space defined by m variables and their covariance. Thus, a small value of Mahalanobis distance increases the chance of an observation to be closer to the group's center and the more likely it is to be assigned to that group. Mahalanobis distance between two samples (**x**, **y**) of a random variable is defined as:

$$\mathbf{d_{mahalanobis}\,(x,y) = \sqrt{}\,(x\text{-}y)^{T}\varepsilon^{-1}(x\text{-}y)} \qquad \mathbf{(10)}$$

ε^{-1} is the inverse co-variance matrix.

In this situation the similarity values lies among the two boundaries, the records are classified as "possible matches" and, in this case, a human judgment is necessary. Usually, most of the existing approaches replica identification depends on several choices to set their parameters, and they may not be always optimal. Setting these parameters requires the accomplishment of the following tasks:

Selecting the best proof to use- as evidence, it takes more time to find out the duplication due to apply more processes to compute the similarity among the data.

Decide how to merge the best evidence—some evidence may be more effective for duplication identification than others. Finding the best boundary values to be used—bad boundaries may increase the number of identification errors (e.g., false positives and false negatives), nullifying the whole process.

To improve the quality of the data in a DBMS is error free and can provide fast outputs. It also concentrates on de-duplication if possible in the data model. The removal of duplicate is not efficient in Government based organization and it is difficult to remove. Avoiding duplicate data provides high retrieval of quality data from huge data set like banking.

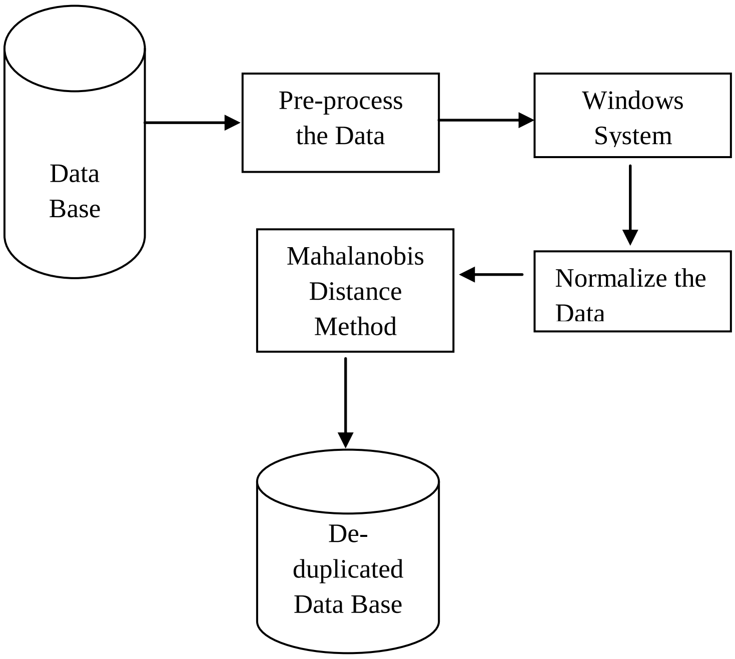


Figure 1: Architecture Diagram

Fig 1 explains how the data should be error free by pre-processing and normalization. Normalized data is divided into sub windows using Windowing system and finally using Mahalanobis distance method the duplicate data is marked and dumped in a separate pool area for further activities. The experiment on the time series data is done in MATLAB software and the time complexity is compared with the existing system. The elapsed time taken for implementing the proposed approach is 5.482168 seconds.

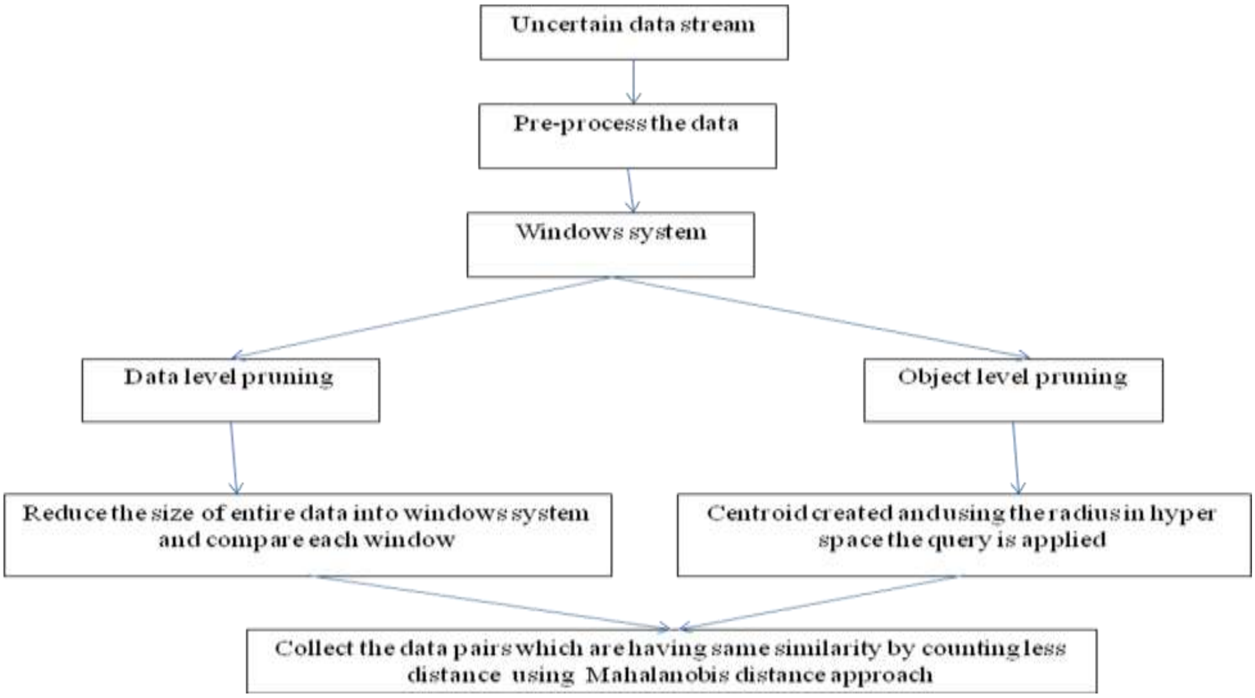

Uncertain data stream
Pre-process the data
Windows system
Data level pruning
Object level pruning
Reduce the size of entire data into windows system and compare each window
Centroid created and using the radius in hyper space the query is applied
Collect the data pairs which are having same similarity by counting less distance using Mahalanobis distance approach

Figure 2 Data Flow Diagram

In fig 2 the uncertain data streams are get pre-processed and the windows system is applied by data level pruning and object level pruning. In data level pruning the data size is reduced by dividing the entire data into windows and each window is compared to find the similarity in the given data. And using object level pruning the centroid get created which and using the radius in hyper space the query is applied. Then collect the data pairs which are having same similarity by counting less distance using Mahalanobis distance approach.

IV PRE-PROCESS DATA

The entire data is read from the time series database containing trading information and investigate that, is there any '~', empty space, "#", "*" and irrelevant characters placed as an entity in the database. If any of these characters presented in an entity location it will be cleared using appropriate function by comparing and investigating the corresponding field [column] data-type. If the data-type of the field is a string then the preprocessing function assigns "NULL" in the corresponding entity else if the data type of the field is a numeric then preprocessing function assigns 0's [according to the length of the numeric data type] in the corresponding entity. Similarly, preprocessing function replaces the entity as today's-date if the data-type is 'date', '*' for data-type is 'character' and so on. This pre-process is done to detect any anomaly; error is in the whole data set.

Data Pre-Processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent and/ or lacking in certain behavior or trends and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. It used database-driven applications such as customer relationship management and rule-based application. Data goes through a series of steps during preprocessing such as: Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. Data Integration: Data with different representations are put together and conflicts within the data are resolved. Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

V WINDOWS SYSTEM (DATA LEVEL PRUNING)

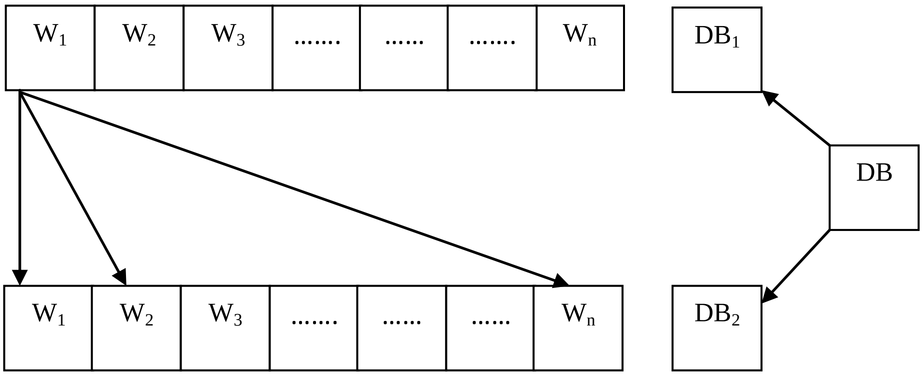


Figure 3 Data Set Divided As Sub-Windows

Fig 3 shows the working of windows system. In windows system, Instead of taking the whole data, it divides the data into sub portions for accurate and faster access. Here, for classification, data level pruning and object level pruning is used. For object level pruning, the data is converted into objects by taking random samples in range. Let the data set is DB and it can be divided as sub windows shown as DB1 and DB2. Each DB1 and DB2 has a number of windows as W_1, W_2… $W_{N.}$ The first window W_1 of DB1 is compared with W_1, W_2… W_N of DB2. Then W_2 of DB1 is compared with W_1, W_2… W_N in DB2. This process takes place up to the windows size is declared. If any error or data redundancy it will be rectified.For any comparison, verification and other relevant tasks the window based data makes easy and fulfill the task very quickly for any DBMS. For example if the data base is having 1000 records can be divided into 4 sub datasets having 25 records each.

Data_ level_ pruning_ Algorithm () the data size is huge in size, in the data level pruning the overall data is divided into multiple sub sets. The complete data DS is divided into DS1 and DS2.

{Data set DS1, DS2 contains set of all data

```
DS1={ W1,W2, …… , WN}
DS2 = {W1, W2… WM}
Enter the value of window W
For I=1 to N step W
For J=1 to M step W
Score[i] = $W_1$ (DB1) – $W_i$ (DB2)
```

```
End j
End i
For i=1 to N
For j=1 to M
If score [I] satisfies {dist (x[i], y[i]) $\leq \varepsilon$} then pair1 = DS1 [I,j], DS2[J,i]
Next j
Next i}
```

Window1 from DB1 is compared with Window1, window2, window3 and so on from DB2 can be written as:

$$\text{Score } [i] = W_1 (DB1) - W_i (DB2) \qquad (11)$$

If the score[i], then both W_1 (DB1) and W_i (DB2) are same and mark it as duplicate. Else W_1 (DB1) is compared with W_{i+1} (DB2).The similarity value among the sub-windows of the dataset DB1 and the dataset DB2 is computed and the result is stored in a variable named score.

$$d[i] = W_1(DB1) - W_2(DB2) \qquad (12)$$

$$\left\{ \begin{array}{l} \text{if } d[i] \leq \alpha \text{ then } 1 \\ \text{if } d[i] \leq 0 \text{ then } 0 \\ \text{if } d[i] > \alpha \text{ then } -1 \end{array} \right\} \qquad (13)$$

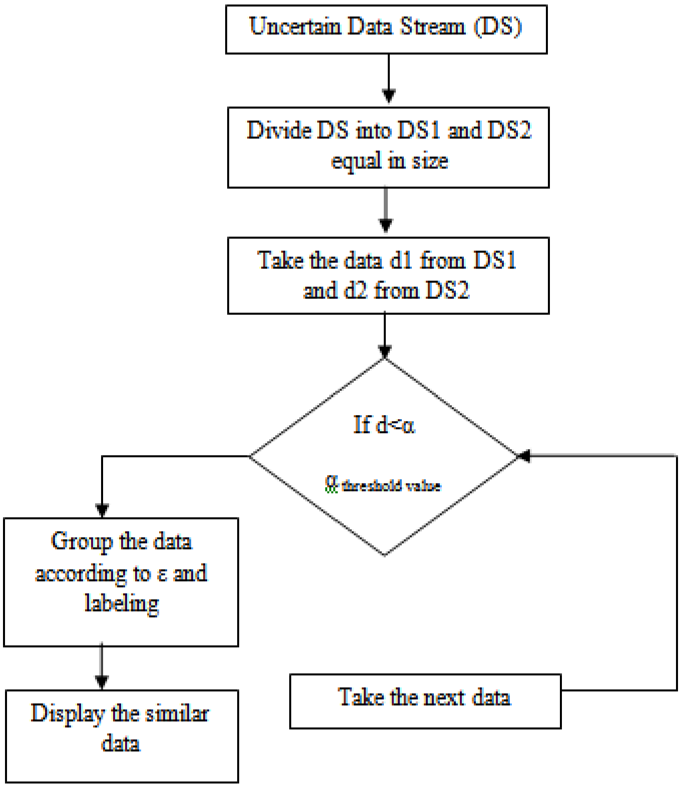

Uncertain Data Stream (DS)
Divide DS into DS1 and DS2 equal in size
Take the data d1 from DS1 and d2 from DS2
If d<α
α threshold value
Group the data according to ε and labeling
Display the similar data
Take the next data

Figure 4 Data Level Pruning System Model

The first line (13) says that the data available in both windows W of DB1 and DB2 are more or less similar, the next line says that exactly same and the third line says that the data are different. Whenever the distance among dataset satisfies d[i] = 0 and d[i] ≤ α both data are marked in the DB. The value of d[i] gives two solutions such as: TPR—if the similarity value lies above this boundary [-1 to 1], the records are considered as replicas; TNR—if the similarity value lies below this boundary, the records are considered as not being replicas

Fig 4 illustrates the data level pruning process which sets the threshold value to group the similar data. DS1 has split into K number of sub windows and the same manner DS2 has split into K number of sub windows. The first element of the first window from DS1 is compared with the second element of the first window and third element of the second window so on. Once the first window of the DS2 is over, the first window of the DS1 is compared with the second window of the DS2, third window of the DS2 and so on. The same manner all the windows from DS1 is compared with the all the windows with the DS2. If d<α data get grouped according to ε and labeling. Then the similar data is displayed. If d is not smaller than threshold value α then take next data and again take some other data to process.

VI OBJECT LEVEL PRUNING

Further the data streams are converted into data objects by taking the random sampling method and it is given in a compartment window which consists of l random sample. The input data is preprocessed, normalized, centroid created and using the radius in hyper sphere, the query is applied. Clearly if all the pair wise distances between samples from two uncertain object X[t+1] and Y[j] are above threshold **ε**, then these two uncertain objects will definitely have their distance above **ε**, and in turn also above $Pr\{\{dist(X[t+1], Y[j]) \leq \varepsilon\}$ = 0. The uncertain data streams converted into hyper sphere data object.

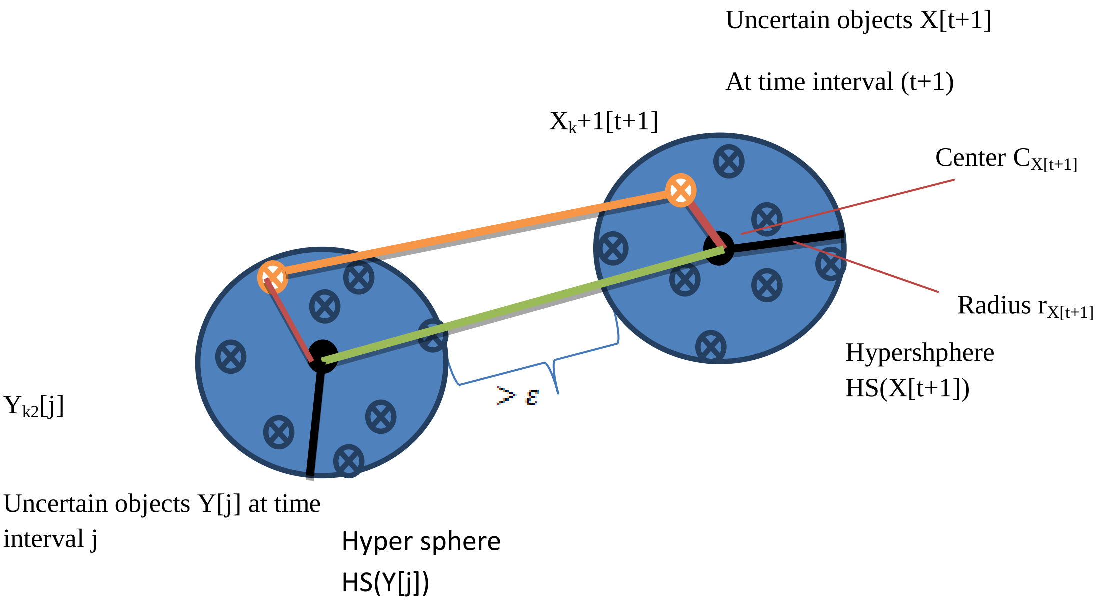


Figure 5 Object-Level Pruning Method

VII NORMALIZE DATA

The time series data which is divided into two sub windows DB1 and DB2 is compared in-between them and error or redundancy in data which is rectified is arranged in an ascending order or descending order. Normalization is the process of efficiently organizing data in a database. There are two goals of normalization (1) eliminating redundant data example storing the same data in more than one table and (2) ensuring data dependencies make same (only storing related data in a table). Both of these are worthy goals as they reduce the amount of space a database consumes and the data is logically stored.

Data in the database can be normalized using any normalization form for fast and accurate query process. In user defined normalization is also applied to improve the efficiency such as arranging the data in a proper manner like ascending order or descending order.

VIII RESULTS AND DISCUSSION

The USG framework and the modified pruning method is experimented and simulated using MATLAB 2012 a software, and the taken for sample is from Benchmark data of US government share market data. The data consists of five field in excel format where the first field says the region, the second field says the data and time, the third filed says the total demand, the fourth field says the RRP and the fifth field says the period type as Trade or non-trade. For this research only the numerical data is taken from excel data and assumed as data stream1 and data stream2.

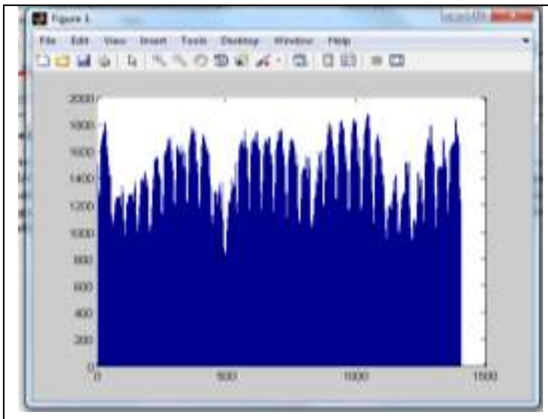
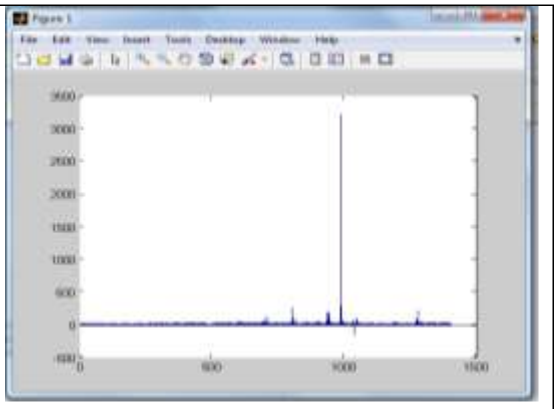
	

Figure-6a: Original uncertain data from Benchmark Database of DS1	Figure-6b: Original uncertain data from Benchmark Database of DS2

The original data of the data stream1 and the data stream2 is shown in the Fig 6a and Fig 6b. There are lot differences in the original data, where the limited size of the data is taken, as 1000 columns is represented as 1x1000 for DS1 and it too same for DS2.

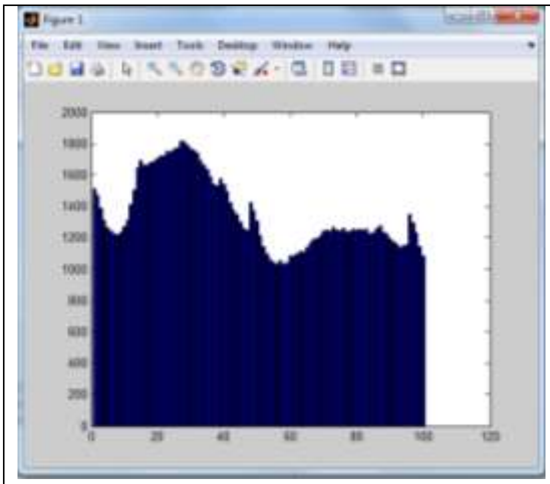


Figure-7a: The Compartment Window for DS1

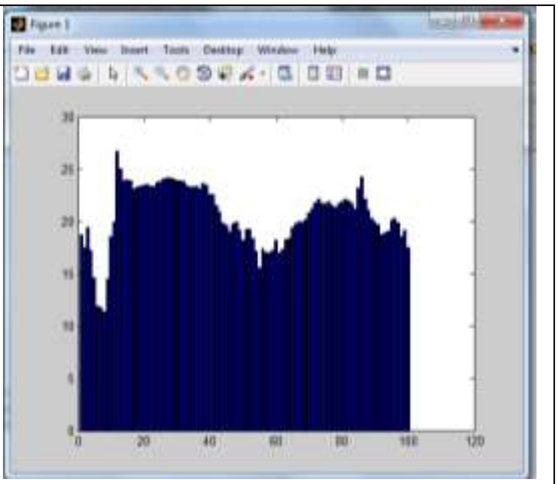


Figure-7b: The Compartment Window for DS2

From the data 1x1000, the compartment window size is 100 and the CW(DS1) and the CW(DS2) is the compartment window data of data stream1 and data stream2 is taken and it is shown in Fig 7a and Fig 7b due to reduce the complexity and fast in preprocessing on the data. The compartment window concept is used sampling for range estimation and it use distance distributions for range estimation and it use reference points for range estimation. So that in this paper also we are using stream joining process the compartment window is used, and the USG will take the most recent CW uncertain data streams as CW(DS1) and CW(DS2).

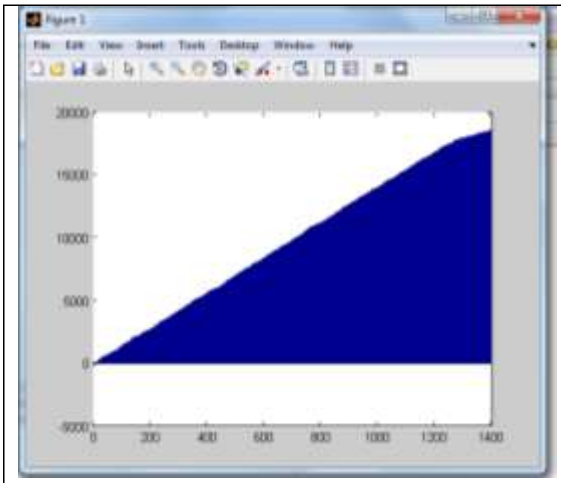


Figure-8a: The Hypersphere data for compartment window of DS1

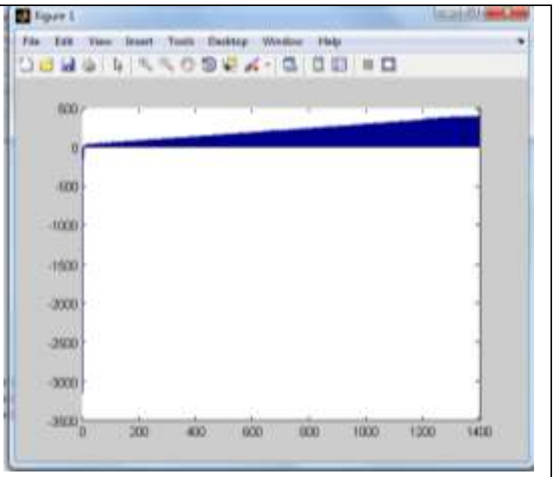
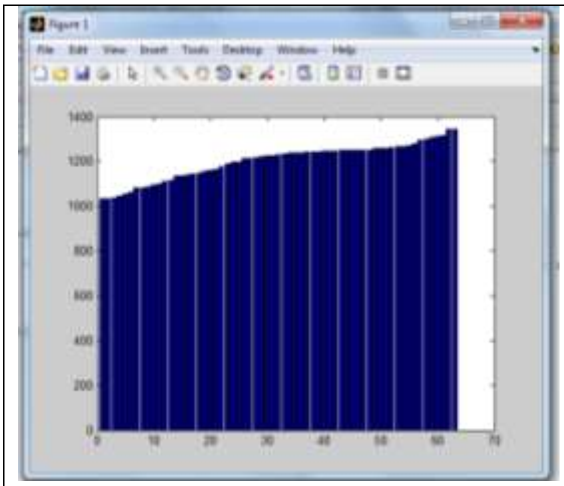
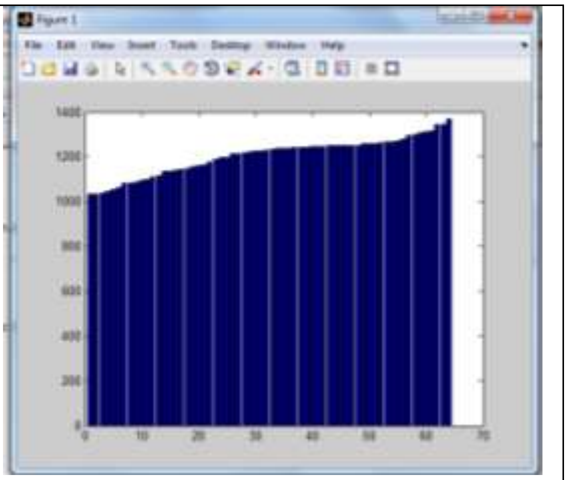


Figure-8b: The Hypersphere data for compartment window of DS1

After selecting the compartment window sized streams the data is converted into hyperspace data and it will sorted for finding the centroid value and the radius. The data converted into hyperspace is by selecting the random sampling for the compartment size as 100. The fig 8a and fig 8b shows the random sampling HS(DS1), HS(DS2) for the CW(DS1) and CW(DS2). Now the comparison of HS(Y[j]) with the HS(X[t=1]) should satisfy the inequality conditions of Equation (7) and (8) will be taken as the matching pairs.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Figure-9a: Data satisfy the In-equal constrains from DS1</td>
<td>Figure-9b: Data satisfy the In-equal constrains from DS2.</td>
</tr>
</table>

The fig 9a and fig 9b gives the data objects from HS(CW(DS1)) and HS(CW(DS2)) which are satisfying the in-equality constraints [5] and [8]. The data objects those who satisfy the unequal condition means we retrieve the original data from the DS1 and DS2 at the time interval t and the pair is denoted as $\langle X[t+1], Y[j] \rangle$.

IX CONCLUSION AND FURTHER WORK

In the proposed work, the Mahalanobis distance approach is taken as the main method for finding the similarity [redundancy] in any database. The time series data is in the form of Excel. De-Duplication eliminates these extra copies by saving just one copy of the data. The simplest form De-Duplication takes place on the file level that is it eliminates duplicate copies of the same file. De-Duplication reduces the amount of disk or tape which is used and it can reduce storage requirement up to 95 percent. It reduces the amount of network bandwidth required for backup process and some cases it can speed up the back up or recovery process. It is concluded from the experiment results obtained using our proposed approach it is easy to do anomaly detection and removal in terms of data redundancy and error. In future the reliability and scalability is investigated in terms of data size and data variations and include so many data sets to find performance analysis.

REFERENCES

[1] Moise´s G. de Carvalho, Alberto H.F. Laender, Marcos Andre´ Gonc¸alves, & Altigran S. da Silva, march 2013, ‟ A Genetic Programming Approach to Record De-duplication ” IEEE transaction on knowledge and data engineering, vol. 24, no.3

[2] Fellegi I.P & Sunter A.B, January 2005, ‟ Duplicate Record Detection: A Survey”, IEEE transaction on knowledge and data engineering, vol. 19, no.1.

[3] Amman P. E, ‟Data Redundancy for the Detection and Tolerance of Software Faults”, 1992, Computing Science and Statistics, vol 45, pp 43-52.

[4] Ahmad Ali Iqbal, Maximilian Ott & Aruna Seneviratne, 2010 ,‟ Removing the Redundancy from Distributed Semantic Web Data”, Database and Expert Systems Applications Lecture Notes in Computer Science, Vol 6261, pp 512-519.

[5] Ireneusz czarnowski & piotr jedrzejowicz, 2010,‟ Active Learning Genetic Programming for Record De-duplication ”, IEEE transactions on knowledge and data engineering vol. 24, no.82.

[6] Elmagarmid A.K & verykios v.s, Jan. 2007, ‟ Duplicate Record Detection For Database Cleansing” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16.

[7] Jose Ramon Cano, Francisco herrera, manuel lozano, 2005, ‟Strategies for Scaling Up Evolutionary Instance Reduction Algorithms for Data Mining”, Book-Soft Computing,Vol 163, pp 21-39.

[8] Nick J. Cercone, Howard J. Hamilton, Xiaohua Hu, Ning Shan, 1997, ‟Data Mining Using Attribute-Oriented Generalization and Information Reduction”, of Rough sets and Data Mning, vol 473, pp 199-22.

[9] Paul Ammann, Dahlard L. Lukes, John C. Knight, 1997, ‟Applying data redundancy to differential equation solvers”, journal of Annals of Software Engineering, Vol 4, Issue 1, pp 65-77.

[10] Peter Christen, September 2012, ‟ A Survey of Indexing Techniques for Scalable Record Linkage and De-duplication” IEEE transactions on knowledge and data engineering, vol. 24 , no. 9.

[11] Rita Aceves-Pérez, Luis Villaseñor-Pineda, Manuel Montes-y-Gomez, 2005, ‟ Towards a Multilingual QA System Based on the Web Data Redundancy”, Computer Science, Vol 3528, pp 32-37.

[12] Yanxu Zhu, Gang Yin, Xiang Li, Huaimin Wang, Dianxi Shi, Lin Yuan, 2011, ‟Exploiting Attribute Redundancy for Web Entity Data Extraction”, Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation Lecture Notes in Computer Science Vol 7008, pp 98-107.

[13] Vikrant Sabnis, Neelu Khare, 2012, ‟An Adaptive Iterative PCA-SVM Based Technique for Dimensionality Reduction to Support Fast Mining of Leukemia Data”,SocProS, vol 23.

[14] Thiago Luís Lopes Siqueira, Cristina Dutra de Aguiar Ciferri, Valéria Cesário Times, Anjolina Grisi de Oliveira, Ricardo Rodrigues Ciferri, June 2009, ‟The impact of spatial data redundancy on SOLAP query performance”,Journal of the Brazilian Computer Society, Vol 15, Issue 2, pp 19-34.