International Journal of Research and Advanced Development (IJRAD)

An Approach to Reconstruct Incomplete Data using DFT and Polynomial Prediction Technique

Ms. S. Kayalvili[1], S. Sabarna[2], R. Sangavi[3], T. SubaPriya[4]

[1]Assistant Professor (Selection Grade), Department of Computer Science & Engineering,
Velalar College of Engineering and Technology, TN, India.
[2, 3, 4]IV Year B. E., CSE, Department of Computer Science & Engineering,
Velalar College of Engineering and Technology, TN, India.
kayalvilis@gmail.com[1], sabarnasubramani@gmail.com[2], rgsangavi@gmail.com[3], subapriya2111@gmail.com[4]

Received Date: 3rd March, 2017, Revised Date: 27th March, 2017, Accepted Date: 9th April, 2017.

Abstract - In a given dataset the returning of k objects that dominating the maximum number of objects is the Top-k Dominating (TDK) query. It plays a major role in decision-making applications by joining the advantages of the skyline and top-k query. There exists many incomplete data in real datasets that are caused by device failure, privacy preservation and so on. In this paper, the systematic study of the top-k query on incomplete data that contains a data having missing dimensional values has implemented. Authors overcome this problem and propose the suite of efficient algorithms for answering TKD queries over incomplete data. The proposed method include novel techniques, such as upper bound score pruning, bitmap pruning, and partial score pruning, to boost query efficiency. The effectiveness of pruning heuristics and the performance of our algorithms will be demonstrated by extensive experimental evaluation by using both real and synthetic datasets.

Keywords – Pruning, k-objects, top-k, Polynomial prediction, Discrete Fourier transform.

I. INTRODUCTION

In this case, partial information arises only when the users enter new data which is unavailable in the database. Every commercial real-world RDBMS will implement some techniques to maintain incomplete data [2], but neither the user nor the RDBMS will not say how the partial information is interpreted. Here now arises a question that does the user of a stock database wanted RDBMS designer to understand the risks and the mission for managing the missing data? Similarly, does epidemiologist gathering information about some disease will think that an RDBMS designer will know how the information about the disease is gathered and why some data is incomplete and what the implications of incomplete data are for his disease and applications? The answer is usually no. When the diversity of the types of incomplete values understood by the database researchers the SQL standard only supports the one type of unmarked value, so RDBMS force users to maintain partial information, in the same way, even some differences, occur. So far we have worked with 2 data sets [5] that hold extensive trial information. One consist information associated with terrifying groups and other is one of the most authoritative data sets. This says about the education of World Bank and UNESCO. This generally contains information for every 221 countries with 4000 attributes per country is being used by educational professionals. The data had been collected manually via extensive surveys, so there are many missing values. The missing values are due to many factors. The most computer scientists seemed to see those missing data very strange as they used to deal with the Amazon and EBay's of the world in which data is collected online and is complete. But unfortunately a huge number of applications where the information is collected through the careful survey and studies. Every year hundreds of millions of dollars have been spent on building the database and maintaining its updating regularly by the World Bank and UNESCO. Not only education policy experts from the World Bank and UNESCO using this data but also the education policy experts from USAID, DFID, AusAid and so on [6]. Because this is simply the most "trusted" data they have. Furthermore, even people from other domains use it.

II. SYSTEM MODELS

A. Existing System

In the existing for handling with incomplete data can be categorized into three such as 1) case deletion, 2) learning without handling of incomplete values, and 3) missing value imputation [7]. For omitting those cases with incomplete values and to use remaining instances to complete learning assignments case deletion is used[8]. The existing technique is simply used to know without maintaining incomplete data by using Bayesian network method and artificial neural network method.

The Parametric and Nonparametric regression imputation methods are the commonly used methods to impute incomplete values [4]. Even that iterative approach impute incomplete values many times that can be developed for

incomplete data imputation. Zhang et al. And the Nonparametric iterative methods based on K-nearest neighbourhood framework was proposed by Caruana.

B. Disadvantages of Existing System

It is mostly impossible to know the distribution of dataset in real applications which hold the parametric and nonparametric regression imputation. So the parametric estimators will lead to highly bias and the optimal control factor could be miscalculated. The imputation methods are developed for either discrete or continuous independent attributes [1]. The methods such as association rule based method and rough set based method are developed to deal with discrete attributes. Some conventional imputation methods developed for discrete attributes using a "frequency estimator".

C. Proposed system

In this, the initial step is pre-processing where the values are removed from the database for certain percentage to produce incomplete data. Secondly, the polynomial prediction is used to find the incomplete values in the database based on upper bound and lower bounds of incomplete value's column. In next step, the values that cannot be predicted by polynomial prediction are identified by Discrete Fourier Transform by some assumption. Finally, RMSE is used to calculate the percentage of error in the predicted values.

The major advantage of polynomial prediction technique is, it is used to predict unknown values with a curve, such as it is based on the time domain. Discrete Fourier Transform is effectively used in online processing. Discrete Fourier Transform will produce an exact result if value remains approximately the same in the frequency domain.

III. WORK DESCRIPTION

A. Data Set Pre-Processing

A data set is a group of data, generally represented in a tabular format. Every column points to a variable and it has various features that define its structure and properties that include the number and its types of variables. The original datasets are deleted with the 5%, 10%, 15%, and 20% of the incomplete values for the selected attribute. Some important tasks in Data Pre-processing include Data Cleaning: Fill in incomplete values, smooth noisy data identify or delete outliers. Data integration: Integration of multiple data cubes, databases or files Data transformation, Normalization, and aggregation. Data reduction: Obtains reduced representation in volume but produces the same or similar analytical results. Data discretization is a part of data reduction preferably for numerical data. Duplicate data sets may have many duplicate values that often look similar. For example, the same person with multiple e-mail addresses. Data Cleaning is the process which deals with the duplications. Discrete Attributes has only a finite or countable finite set of data.

B. Polynomial Prediction

The polynomial prediction is based on replacing the missing values with their corresponding predicted values from fitting polynomial model plus a random error [3]. This procedure was used for incomplete longitudinal data. However, this modification for missing data in polynomial wavelet regression should be considered. The steps of this procedure are illustrated as follows:

1) In its place of deleting that involves any missing as in literature, missing data is filled using the non-parametric process.
2) Find the ordinary least squares fit of a quadratic polynomial. Choose the 2nd order models in it are the lowest degree polynomial that admits non-zero coefficients.
3) Find the predicted values according to the complete data set.
4) And calculate RMSE.
5) Replace the missing values Here MSE refers to the mean squared error from the 2nd order polynomial model.
6) Iterate the last two steps for r times (say r=100) or until convergence.
7) Finally construct the complete data Y = (y1,y2,...,yn)

C. Non-Parametric Process

Non-parametric process in the sense of a statistic over data, which is defined as a function on a sample that has no dependence on a parameter. The foremost process in Polynomial Prediction is the non-parametric process to predict the missing values. If the value is missed for a record in a specific field, it not only refers the base table but also refers the supporting tables to collect the historical data to predict the missing values. The reference is done with the help of primary key that relates the tables. Historical data is collected through the non-parametric process. The important step in Polynomial Prediction is curved analysis. PP involves two steps to predict the missing values in the table. In the first step it analysis the historical data collected in the non-parametric process and approximates the historical data. The deviation in the polynomial curve indicates the existing of missing value. In the next step, it predicts the missing values according to the curves.

It finds the predicted values according to the complete data set. Second-order model is chosen as it is the lowest degree polynomial that admits non-zero coefficients.

D. DFT Prediction

The algorithm consists of four steps:

1) Divide it into h disjoint segments with identical length l,
2) Obtain the mean value avg for each segment i, where 1 <i<h,
3) Convert each missing value into a symbol si, that is, transforming the subsequence to its symbolic representation s1 s2 . . . sh, and
4) Insert the string s1 s2 . . . sh into the aggregate with height h.

The first three steps transform the time series to its symbolic representation. Assume that the value domain of a time series T0 is [-1.5, 1.5]. We partition it into three smaller ranges of equal size, say, [0.5, 1.5], [-0.5, 0.5], and [-1.5, -0.5], it corresponds to a, b, and c symbols, respectively. Note that in case a prior knowledge of stream data is known then one can divide the value domain into small ranges of different lengths. As a second step, we divide the time series T0 of length H into seven segments of equal length, take the average value avg within each segment i, and finally convert each avgi into a unique symbol.

After the discretization, T0 is represented by a string consisting of ordered symbols. Specifically, the discrete version of T0 is the string "ababbaa". Space efficiency is one of the advantages in discrete time series. That is, if there are in total K symbols in the alphabet table, each sequence requires only limited bits at most.

The fourth step inserts all the time series strings into an aggregate string. However, in contrast to an ordinary string, each node entry in our aggregate string contains a triple < freq; hit; miss >, where freq is the frequency in the time series that a string appears, hit is the times that our prediction succeeds, and miss is the times that it fails. Intuitively, if the frequency freq is high, then it will have a higher probability of repeating again. Or else, if our prediction fails quite often for a certain string, that is, its aggregate miss is large, and then we have to lower the chance of choosing that string as the result of the prediction.

E. Computation of RMSE

The Root-Mean-Square Error (RMSE) is often used as a measure of the differences between values predicted by a model and the values noted. These residuals perform calculations over the data sample that was used for estimation and are called prediction errors when they are computed out-of-sample. The RMSD combine the magnitudes of the errors in predictions into a single measure from various times of predictive power. RMSD is a good measure of accuracy, but only to evaluate forecasting errors of unlike models for a particular variable and not between other variables, because it is scale-dependent.

The RMSE is used to assess the predictive ability after the algorithm has converged:

RMSE - $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i - e_i')^2}$

Where e_i - is the original attribute value,

e_i' - is the estimated attribute value, and

m - is the total number of predictions.

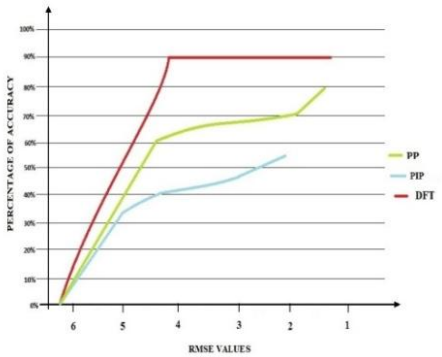

Fig. 1 Comparison Chart

The larger the value of the RMSE, the prediction is less accurate. At that time, the correlation coefficient between the actual and after convergence the predicted values of missing attributes is calculated. Figure 1 shows the comparison of

existing and proposed algorithm. Experimental result shows that, the proposed system is more accurate with a bonus of 30% more accuracy.

IV. CONCLUSION

PP and DFT based estimators have proposed next to the case that data sets have both continuous and discrete independent attributes. It utilizes all available observed information, including incomplete instances, to impute missing values, whereas existing imputation methods use only for complete instances. The optimal bandwidth is experimentally selected by this method. This experiment result has demonstrated that the proposed algorithms outperform the existing ones for imputing both discrete and continuous missing values.

V. FUTURE WORK

The future work consists of the following sequence; initially, a nonparametric iterative imputation method is presented. In this, functions for the discrete attributes are studied and then a mixture function is proposed by combining a discrete data and continuous data. Also the chronicle data should also take into account to make a resourceful retrieval.

REFERENCES

[1] Ahamad I.A. and P.B. Cerrito, "Nonparametric Estimation of Joint Discrete-Continuous Probability Densities with Applications,"J. Statistical Planning and Inference, vol. 41, pp. 349-364, 1994.
[2] Allison P., Missing Data. Sage Publication, Inc., 2001.
[3] Batista G. and M. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," Applied Artificial Intelligence, vol. 17, pp. 519-533, 2003.
[4] M. L.Brown, "Data Mining and the Impact of Missing Data," Industrial Management and Data Systems, vol. 103, no. 8, pp. 611-621, 2003.
[5] Cristian Molinaro, Maria Vanina Martinez, John Grant, and V. S. Subrahmanian, "Customized Policies for Handling Partial Information in Relational Databases," Knowledge and Data Engineering,2012.
[6] N. Pal, L. Jain, and N. Teoderesku, "Information Systems - Trends in Data Mining and Knowledge Discovery", Springer, 2002.
[7] J. Han and M. Kamber, Data Mining Concepts and Techniques, second ed. Morgan Kaufmann Publishers, 2006.
[8] Lakshminarayan K., "Imputation of Missing Data in Industrial Databases," Applied Intelligence, vol. 11, pp. 259-275, 1999.