National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications (NCACSA 2012)
Proceedings published in International Journal of Computer Applications® (IJCA)

An Efficient Approach for Filling Incomplete Data

P.M.Kiran , A.Prakash Rao , B.Ratnamala
Assistant Prof.- Gayatri Vidya Parishad College of Engineering, Department of Computer Applications, Visakhapatnam, India

ABSTRACT
Good data preparation is a key prerequisite to successful data mining. Conventional wisdom suggests that data preparation takes about 60 to 80% of the time involved in a data mining exercise. There have been good reviews of the problems associated with data preparation. However the data preprocessing is a crucial step used for variety of data warehousing and mining. Real world data is noisy and can often suffer from corruptions or incomplete values that may impact the models created from the data. Accuracy of any mining algorithm greatly depends on the input datasets. In this paper we describe a novel idea of predicting the missing values in the dataset by a well known principle of Maximum likelihood EM (Expectation Maximization). After doing implementing and applying the EM filter, the dataset is completed with the estimated values, based on the well known principle of expected maximization of attribute instance. We demonstrate the efficacy of the approach on real data sets as a preprocessing step.

Keywords
Data mining, Data preprocessing, Missing data.

1. INTRODUCTION
Many data analysis applications such as data mining, web mining, and information retrieval system require various forms of data preparation. Mostly all this worked on the assumption that the data they worked is complete in nature, but that is not true! [1]

In data preparation, one takes the data in its raw form, removes as much as noise, Redundancy and incompleteness as possible and brings out that core for further processing[1]. Common solutions to missing data problem include the use of imputation, statistical or regression based procedure[7].

We note that, the missing data mechanism would rely on the fact that the attributes in a data set are not independent from one another, but that there is some predictive value from one attribute to another. [3]

So far we have several approaches to fill missing values in a dataset. Among that some of the traditional techniques that are frequently used are:

a). Desired value technique

b). Mean value technique

In desired value technique the missing values are filled by any random data in the attribute. And in the mean value technique the missing values are filled with the mean of the attribute. These two techniques are not so efficient and the resultant dataset after filling data may not be exact as original data.[9]

Therefore we used the well known machine learning estimation technique, Expectation Maximization i.e. EM [7], for predicting the missing values.

EM is a new technique for predicting the missing values and it is a general approach to iterative computation of maximum likelihood estimates when a observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step followed by a maximization step we call it the EM algorithm. The EM process is remarkable in part because of the simplicity and generality of the associated theory, and in part because of the wide range of examples which fall under its umbrella. When the underlying complete data come from an exponential family whose maximum-likelihood estimates are easily computed, then each maximization step of an EM algorithm is likewise easily computed.[2]

The term "incomplete data" in its general form implies the existence of two sample spaces %Y and X and a many-one mapping from3 to Y. The observed data y are a realization from *CY.*

The corresponding x in X is not observed directly, but only indirectly through y. More specifically, we assume there is a mapping x+ y(x) from X to Y, and that x is known only to lie in X(y), the subset of X determined by the equation y = y(x), where y is the observed data.

We refer to x as the ***complete data*** even though in certain examples x includes what are traditionally called parameters. We postulate a family of sampling densities f(xI +) depending on parameters and derive its corresponding family of sampling densities g(y1 +). The complete-data specification ***f(***...1 ...) is related to the incomplete-data specification g(...I ...).

The EM algorithm is directed at finding a value of + which maximizes g(y1 +) given an observed y, but it does so by making essential use of the associated family f(xl +). Notice that given the incomplete-data specification g(y1 +), there are many possible complete-data specifications f (x1+) that will generate g(y1 +). Sometimes a natural choice will be obvious, at other times there may be several different ways of defining the associated f(xl +).

Each iteration of the EM algorithm involves two steps which we call the expectation step (E-step) and the maximization step (M-step). The precise definitions of these steps, and their associated heuristic interpretations, are given. [2]

Here we shall present only a simple example to give the flavor of the method.

Consider two coins A and B. If two coins are tossed as show in the figure 1 i.e., coins are tossed 10 times per set, there may be chance of head or tail for each toss so probability for head is ½ and similarly for tail is ½ . Here in this example we are considering 5 sets. For the first set coin B is tossed, the occurrence of head and tail are as shown in the fig 1. For the second set coin A is tossed and the occurrence of head and tails are as shown in fig 1.In the fig 1 the set in blue color indicates coin B and set with red color indicates

coin A. The maximum likelihood θ of coin A and Coin B are calculated as show in the fig1.

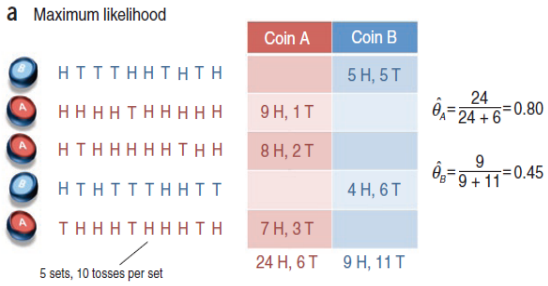

Fig 1: Maximum likelihood of two coins

After calculating the maximum-likelihood of two coins the next step is expectation maximization calculation which is as shown in fig 2.

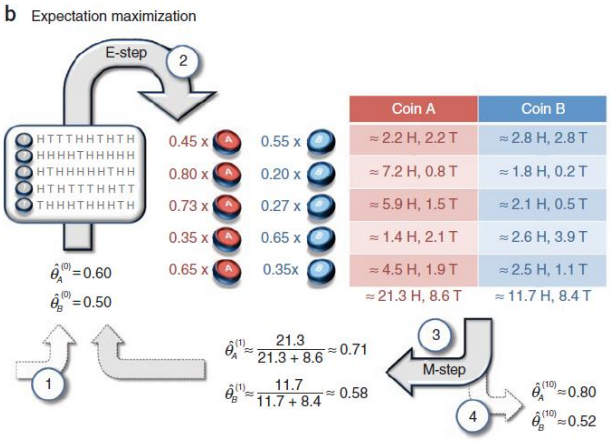

Fig 2: Calculation of expectation

2. ALGORITHM

This algorithm is designed to give the user an understanding of the EM algorithm. EM is a common technique for finding missing values to:

i) Predict missing values by most probable estimated values.
ii) Estimate parameters.
iii) Re-estimate the missing values assuming the new parameter estimates are correct.
iv) Re-estimate parameters, and so forth, iterating until convergence.

An Expectation-Maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated.

Algorithm: Maximum-Likelihood Expectation Maximization: A New Approach For Missing Value Prediction
Steps:

1: Take the Input datset as ARFF format.

2: Replace the missing value by zero.

3: Repeat Step 4 to Step 10 until all columns are called.

4: Take the probabilities as

(ai+bi θ)… (an +bn θ) for each occurrence.

5: Calculate θhat=d/dθ(log(Ai(ai+biθ+…..Ai(ai+biθ)))

$F(\theta) = Ai_{r-1}(b_{r-1}\,\theta)/(a_r\,\theta + b_r\,\theta)$

6: E step:

$Ai_r = Ai_{r-1}(F(\theta))$.

7: M-step

$\theta_{t+1} = \theta_{hat}$ at t^{th} iteration.

8: Convergence Step

Take $\theta_t = 0.5$

Repeat till $(\theta_{t+1} - \theta_{hat}/\ \theta_t - \theta_{hat}) = (\theta_{t+2} - \theta_{hat}/\ \theta_{t+1} - \theta_{hat})$

9: At convergence, get actual θ_{hat}.

10**:** a) After completion of one iteration, check next missing Instance.

b) Repeat the procedure

Table 1: Comparison of three approaches for filling missing values

Original Dataset	Datasewt with Missing values	Dataset with Desired value Technique	Dataset with Mean Value Technique	Dataset with EM Technique
Original dataset CPU Dataset with 7 missing values Output after applying filter @relation 'cpu' @attribute MYCT real @attribute MMIN real @attribute MMAX real @attribute CACH real @attribute CHMIN real @attribute CHMAX real @attribute c lass real @data 125, 256, 6000, 256,16, 128,199 29, 8000,32000,32, 8, 32, 253 29, 8000,32000,32, 8, 32, 253 29, 8000,32000,32, 8, 32, 253 29, 8000,16000,32, 8, 16, 132 26, 8000,32000,64, 8, 32, 290 23, 16000,32000,64, 16, 32, 381 23, 16000,32000,64, 16, 32, 381 23, 16000,64000,64, 16, 32, 749 23, 32000,64000,128,32 , 64, 1238	@relation 'cpu' @attribute MYCT real @attribute MMIN real @attribute MMAX real @attribute CACH real @attribute CHMIN real @attribute CHMAX real @attribute class real @data 125, 256, 6000, 256, 16, 128, 199 29, 8000,32000, 32, 8, 32, 253 29, **?,** 32000, 32, 8, **?**, 253 29, 8000,32000, 32, 8, 32, **?** 29, 8000, **?,** 32, 8, 16, 132 26, 8000,32000, 64, 8, 32, 290 **?,** 16000,32000, **?,** 16, 32, 381 23, 16000,32000, 64, **?** , 32, 381 23, 16000, 64000, 64, 16, 32, 749 23, 32000,64000,128 , 32, 64, 1238	@relation 'cpu' @attribute MYCT real @attribute MMIN real @attribute MMAX real @attribute CACH real @attribute CHMIN real @attribute CHMAX real @attribute class real @data 125,256,6000,256,16, 128,199, 29,8000,32000,32,8,32, 253 29,16531.20,32000,32,8, 94.56,253 29,8000,32000,32,8,32, 1164.05 29,8000,22613.27,32,8, 16,132 26,8000,32000,64,8,32, 290 13.64,16000,32000,136.07, 16,32,381 23,16000,32000,64,23.99,32, 381 23,16000,64000,64,16 32,749 23,32000,64000,128,32 64,1238	@relation 'cpu' @attribute MYCT real @attribute MMIN real @attribute MMAX real @attribute CACH real @attribute CHMIN real @attribute CHMAX real @attribute class real @data 125,256,6000,256, 16,128,199 29,8000,32000,32, 8,32,253 29,11225.6,32000,32, 8,40,253 29,8000,32000,32,8,32 387.6 29,8000,32600,32,8, 16,132 26,8000,32000,64,8, 32,290 33.6,16000,32000,70.4,16, 32,381 23,16000,32000,64,12, 32,381 23,16000,64000,64,16, 32,749 23,32000,64000,128,32, 64,1238	@relation cpuweka. filters.unsupervised.a ttribute.EM @attribute MYCT numeric @attribute MMIN numeric @attribute MMAX numeric @attribute CACH numeric @attribute CHMIN numeric @attribute CHMAX numeric @attribute class numeric @data 125, 256, 6000, 256, 16, 128, 199 29, 8000, 32000, 32, 8, 32, 253 29, 8001.02, 32000, 32, 8, 33.00, 253 29, 8000, 32000, 32, 8, 32, 254.00 29, 8000, 32001.04 ,32, 8, 16, 132 26, 8000, 32000, 64, 8, 32, 290 27.42,16000, 32000, 65.00,16, 32, 381 23, 16000, 32000, 64, 17.00, 32, 381 23, 16000, 64000, 64, 16, 32, 749 23, 32000, 64000, 128, 32, 64, 1238

3. EXPERIMENTAL RESULTS

3.1 Approach

The objective of our experiment is to build the filter as a preprocessing step in Weka Workbench, which completes the data sets from missing data sets.

We did not intentionally select those data sets in UCI [8], which originally come with missing values because even if they do contain missing values, we don't know the accuracy of our approach. For experimental set up, we take the complete dataset from UCI repository [8], and then deliberately deleted some values for making it as an incomplete datasets.

3.2 Results

In Table 1, we used the UCI [8] dataset CPU, in the original dataset, there are seven numeric attributes. The first column of Table 1 gives the original dataset values. In the second column of Table 1gives the dataset with missing values. In the third and fourth columns of Table 1 we purposely deleted seven values using desired value technique and Mean value technique for making it incomplete datasets. Finally in the last column, after applying the EM filter, we get the estimated values. These estimated values as compared to the original values are in the same domain, therefore gives the expected results.

The graphical representation of original dataset with seven missing values and the dataset that occurred after applying desired value technique for missing values is as shown in fig 3.In the graph blue line indicates original dataset and red line indicates dataset after applying desired value technique.

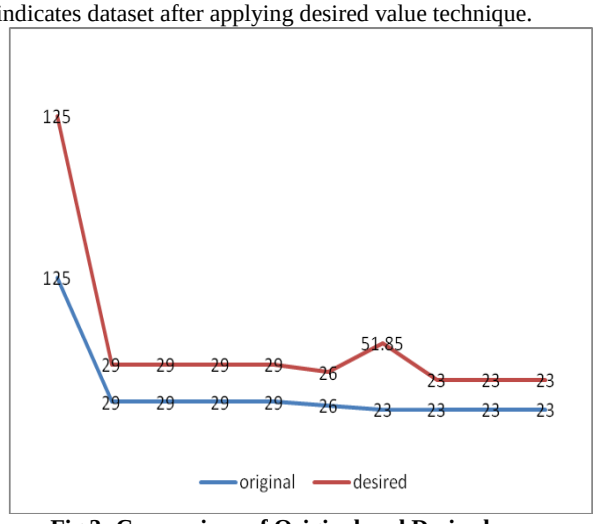

Fig 3: Comparison of Original and Desired

The graphical representation of original dataset with seven missing values and the dataset that occurred after applying mean value technique for missing values is as shown in fig 4.In the graph blue line indicates original dataset and red line indicates dataset after applying mean value technique.

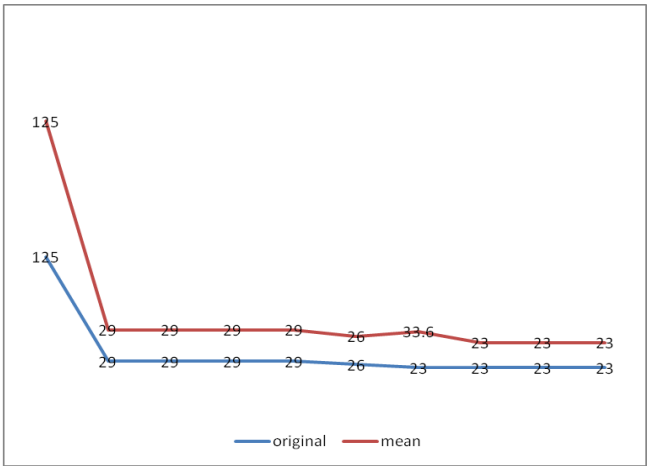

Fig 4: Comparison of Original and Mean

The graphical representation of original dataset with seven missing values and the dataset that occurred after applying Maximum-likelihood: EM algorithm for missing values is as shown in fig 5.

In the graph blue line indicates original dataset and red line indicates dataset after applying EM algorithm.

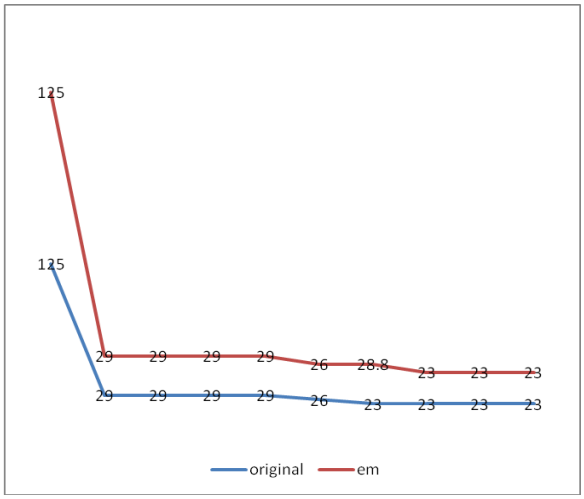

Fig 5: Comparison of Original and EM

After observing the three graphs, the proposed EM algorithm that fills the missing values in a dataset by considering maximum-likelihood of each attribute is an effective method for predicting missing values.

4. CONCLUSION

Expectation Maximization is used to recommend incomplete instances in a dataset for information completion, where attribute of instances mixing the missing information of different attributes are inheritably different and data is bounded by specific budget.

The design of our Algorithm distinguishes our work from existing approaches including the efficiency with the previous approaches and the predicted value using EM algorithm is found to be either lying very close to real value or show an attribute relation.

5. REFERENCES

[1] Sameer S. Prabhune, Dr. S.R. Sathe "Reconstruction of a Complete Dataset from an IncompleteDataset by Expectation Maximization Technique", International

Journal of Computer Science and Network Security, VOL.10 No.11, November 2010

[2] Data Preparation for Data Mining, D Pyle, 1999, Morgan Kaufmann Inc.,ISBN 1-55860-529-0.

[3] S.Parthsarthy and C.C. Aggarwal, "On the Use of Conceptual Reconstruction for Mining Massively Incomplete Data Sets, "IEEE Trans. Knowledge and Data Eng., pp. 1512-1521,2003.

[4] J. Quinlan, C4.5: Programs for Machine Learning, San Mateo, Calif.: Morgan Kaufmann, 1993.

[5] S. Mehta,S.Parthsarthy and H. Yang " Toward Unsupervised correlation preserving discretization", IEEE Trans. Knowledge and Data Eng.,pp 1174- 1185 ,2005.

[6] Ian H. Witten and Eibe Frank , "Data Mining: Practical Machine Learning Tools and Techniques" Second Edition, Morgan Kaufmann Publishers. ISBN:81-312-0050-

[7] R. Little, D. Rubin. Statastical Analysis with Missing Data. Ch.8 , pp 164-172,Wiley Series in Probability and Statistics, 2002.

[8] UCI Machine Learning Repository, [9] Jiawei Han and Micheline Kamber "Data Mining Concepts and techniques "

[10] M.richardson and P.Domingos. Mining Knowledge – sharing sites for viral marketing.

[11] Data Mining Leading Edge: Insurance & Banking, D Romano in *Proceedings of Knowledge Discovery and Data Mining*, Unicom, Brunel University, 1997.

[12] Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem,M A Hernandez and S J Stolfo, Data Mining and Knowledge Discovery 2,p1-31, 1998.