International Journal of Data Engineering (IJDE)

ISSN : 2180-1274

Volume 1, Issue 2

Number of issues per year: 6

International Journal of Data Engineering (IJDE)

Volume 1, Issue 2, 2010

Edited By

Computer Science Journals

www.cscjournals.org

International Journal of Data Engineering (IJDE)

Book: 2010 Volume 1, Issue 2
Publishing Date: 31-07-2010
Proceedings
ISSN (Online): 2180 -1274

IJDE Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is Second issue of volume one of the International Journal of Data Engineering (IJDE). IJDE is an International refereed journal for publication of current research in Data Engineering technologies. IJDE publishes research papers dealing primarily with the technological aspects of Data Engineering in new and emerging technologies. Publications of IJDE are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJDE is Annotation and Data Curation, Data Engineering, Data Mining and Knowledge Discovery, Query Processing in Databases and Semantic Web etc.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJDE is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJDE as one of the top International journal in Data Engineering, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Data Engineering fields.

IJDE editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJDE provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Data Engineering (IJDE)

Editorial Board

Table of Content

Volume 1, Issue 2, July 2010.

Pages

Some Imputation Methods to Treat Missing Values in Knowledge Discovery in Data warehouse

D. Shukla diwakarshukla@rediffmail.com
Deptt. of Mathematics and Statistics,
Dr. H.S.G. Central University, Sagar (M.P.), India.

Rahul Singhai singhai_rahul@hotmail.com
Iinternational Institute of Professional Studies,
Devi Ahilya Vishwavidyalaya, Indore (M.P.) India.

Narendra Singh Thakur nst_stats@yahoo.co.in

B.T. Institute of Research and Technology,
Sironja, Sagar (M.P.) India.

Naresh Dembla nareshdembla@gmail.com
Iinternational Institute of Professional Studies,
Devi Ahilya Vishwavidyalaya, Indore (M.P.) India.

Abstract

One major problem in the data cleaning & data reduction step of KDD process is the presence of missing values in attributes. Many of analysis task have to deal with missing values and have developed several treatments to guess them. One of the most common method to replace the missing values is the mean method of imputation. In this paper we suggested a new imputation method by combining factor type and compromised imputation method, using two-phase sampling scheme and by using this method we impute the missing values of a target attribute in a data warehouse. Our simulation study shows that the estimator of mean from this method is found more efficient than compare to other.

Keywords: KDD (Knowledge Discovery in Databases), Data mining, Attribute, Missing values, Imputation methods, Sampling.

1. INTRODUCTION

"Data mining", often also referred to as "Knowledge Discovery in Databases" (KDD), is a young sub-discipline of computer science aiming at the automatic interpretation of large datasets. The classic definition of knowledge discovery by Fayyad et al.(1996) describes KDD as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al. 1996). Additionally, they define data mining as "a step in the KDD process consisting of applying data analysis and discovery algorithms. In order to be able to "identify valid, novel patterns in data", a step of pre-processing of the data is almost always required. This preprocessing has a significant impact on the runtime and on the results of the subsequent data mining algorithm.

The knowledge discovery in database is more than pure pattern recognition, Data miners do not simply analyze data, and they have to bring the data in a format and state that allows for this

analysis. It has been estimated that the actual mining of data only makes up 10% of the time required for the complete knowledge discovery process (Pyle 1999). In our opinion, the precedent time-consuming step of preprocessing is of essential importance for data mining (Han and Kamber 2001). It is more than a tedious necessity: The techniques used in the preprocessing step can deeply influence the results of the following step, the actual application of a data mining algorithm (Hans et al.(2007). We therefore feel that the role of the impact on and the link of data preprocessing to data mining will gain steadily more interest over the coming years.

Thus Data pre-processing is one of the essential issue of KDD process in Data mining. Since data warehouse is a large database that contains data that is collected and integrated from multiple heterogeneous data sources. This may lead to irrelevant, noisy inconsistent, missing and vague data. So it is required to apply different data pre-processing techniques to improve the quality of patterns mined by data mining techniques. The data mining pre-processing methods are organised into four categories: Data cleaning, data integration and transportation, data reduction, descritization and concept hierarchy generation.

Since the goal of knowledge discovery can be vaguely characterized as locating interesting regularities from large databases (Fayyad et al. &. Krishnamurthy R. et al.) For large collections of data, sampling is a promising method for knowledge discovery: instead of doing complicated discovery processes on all the data, one first takes a small sample, finds the regularities in it, and then possibly validates these on the whole data

Sampling is a powerful data reduction technique that has been applied to a variety of problems in database systems. Kivinen and Mannila (1994) discuss the general applicability of sampling to data mining, and Zaki, et al.(1996) employ a simple random sample to identify association rules. Toivonen (1996) uses sampling to generate candidate itemsets but still requires a full database scan. John and Langley (1996) give a dynamic sampling method that selects the sample size based on the observed behavior of the data-mining algorithm. Traditionally, random sampling is the most widely utilized sampling strategy for data mining applications. According to the Chernoff bounds, the consistency between the population proportion and the sample proportion of a measured pattern can be probabilistically guaranteed when the sample size is large (Domingo et al.(2002) and Zaki et al.(1997)). Kun-Ta Chuang et al.(2007) proposed a novel sampling algorithm (PAS) to generate a high quality online sample with the desired sample rate.

Presence of missing data is one of the critical problem in data cleaning and data reduction approach. While using sampling techniques to obtain reduced representation of large database, it often possible that the sample may contains some missing values.Missing data are a part of most of the research, and missing data can seriously affect research results (Robert 1996). So, it has to be decided how to deal with it. If one ignores missing data or assumes that excluding missing data is acceptable, there is a risk of reaching invalid and non-representative conclusions. There are a number of alternative ways of dealing with missing data (Joop 1999). There are many methods of imputation (Litte and Rubin 1987) like Mean Imputation,regression imputation, Expectation maximization etc. Imputation of missing data minimizes bias and allows for analysis using a reduced dataset. In general the imputation methods can be classified into single & multiple imputations. The single imputation method always imputes the same value, thereby ignoring the variance associated with the imputation process. The multiple imputations method imputes several imputed values and the effect of the chosen imputed values on the variance can be taken into account.

Both the single-imputation and MI methods can be divided into three categories: 1) data driven; 2) model based; and 3) ML based (Laxminarayan et al.(1999), Little and Rubin(1987), Oh (1983)). Data-driven methods use only the complete data to compute imputed values. Model-based methods use some data models to compute imputed values. They assume that the data are generated by a model governed by unknown parameters. Finally, ML-based methods use the entire available data and consider some ML algorithm to perform imputation. The data-driven methods include simple imputation procedures such as mean, conditional mean, hot-deck, cold-deck, and substitution imputation (Laxminarayan et al. (1999), Sarle(1998)). Several model-based imputation algorithms are described by Little and Rubin (1987). The leading methods include regression-based, likelihood-based, and linear discriminant analysis (LDA)-based imputation. In regression-based methods, missing values for a given record are imputed by a regression model based on complete values of attributes for that record. The likelihood-based methods can be

considered to impute values only for discrete attributes. They assume that the data are described by a parameterized model, where parameters are estimated by maximum likelihood or maximum a posteriori procedures, which use different variants of the EM algorithm (Cios(1998), Little and Rubin(1987)). A probabilistic imputation method that uses probability density estimates and Bayesian approach was applied as a preprocessing step for an independent module analysis system (Chan K et al.(2003)). Neural networks were used to implement missing data imputation methods (Freund and Schapire (1996), Tresp (1995)). An association rule algorithm, which belongs to the category of algorithms encountered in data mining, was used to perform MIs of discrete data (Zhang (2000)). Recently, algorithms of supervised ML were used to implement imputation. In this case, imputation is performed one attribute at a time, where the selected attribute is used as a class attribute. Several different families of supervised ML algorithms, such as decision trees, probabilistic, and decision rules (Cios et al.(1998)) can be used; however, the underlying methodology remains the same. For example, a decision tree C4.5 (Quinlan(1992),(1986), and a probabilistic algorithm A decision rule algorithm CLIP4 (Cios(1998)) and a probabilistic algorithm Naïve Bayes were studied in (Farhangfar et al.(2004). A k-nearest neighbor algorithm was used by Batista and Monard(2003). Backpropagation Neural Network (BPNN) is one of the most popular neural network learning algorithms. Werbos (1974) proposed the learning algorithm of the hidden layers and applied to the prediction in the economy. Classification is another important technique in data mining. A decision tree approach to classification problems were described by Friedman 1997. Let $A=\{x,y,z....\}$ is a finite attribute set of any database, where target attribute domain Y consist of $Y_i;(i=1,2,.....N)$ values of main interest and attribute domain X consist of $X_i;(i=1,2,.....N)$ auxiliary values, that is highly associated with attribute domain Y. Suppose target attribute Domain Y has some missing values. Let $\overline{y}$ be the mean of finite attribute set Y under consideration for estimation $\left[\overline{Y}=N^{-1}\sum_{i=1}^{N}Y_i\right]$ and $\overline{X}$ be the mean of reference attribute set X. When $\overline{X}$ is unknown, the two-phase sampling is used to estimate the main data set missing values (Shukla, 2002).

2. PROPOSED IMPUTATION TECHNIQUES FOR MISSING ATTRIBUTE VALUES

Consider preliminary large sample $S'=\{X_i;i=1,2,3,.....,n'\}$ of size n' drawn from attribute data set A by SRSWOR and a secondary sample of size n $(n<n')$ drawn in the following manner (fig. 1).

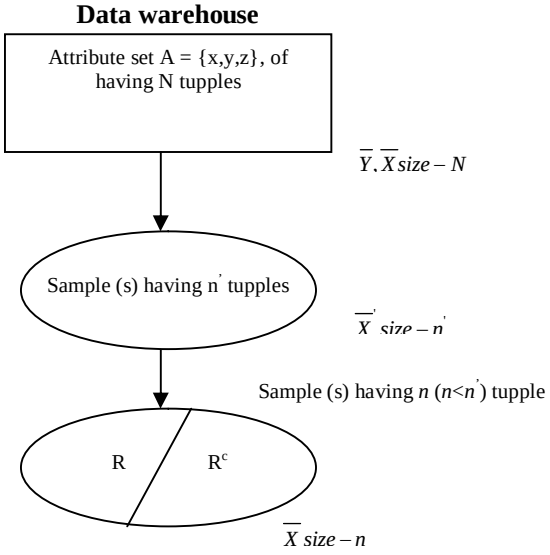

FIGURE 1.

The sample S of n units contains r available values (r < n) forming a subspace R and (n – r) missing values with subspace R^C in $S = R \cup R^C$. For every $i \in R$, the y_i's are available values of attribute Y and for $i \in R^C$, the y_i values are missing and imputed values are to be derived, to replace these missing values.

2.1.0 F-T-C Imputation Strategies:

For $y_{ji}\,(j = 1,2,3)$

$$y_{ji} = \begin{cases} \left(\frac{kn}{r}\right) y_i + (1-k)\phi_j'(k) & if\ i \in R \\ (1-k)\phi_j'(k) & if \quad i \in R^C \end{cases} \qquad \ldots(2.1)$$

where, $\phi_1'(k) = \bar{y}_r \left[\frac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}}\right]$; $\phi_2'(k) = \bar{y}_r \left[\frac{(A+C)\bar{x} + fB\bar{x}_r}{(A+fB)\bar{x} + C\bar{x}_r}\right]$;

$\phi_3'(k) = \bar{y}_r \left[\frac{(A+C)\bar{x}' + fB\bar{x}_r}{(A+fB)\bar{x}' + C\bar{x}_r}\right]$; $A = (k-1)(k-2)$; $B = (k-1)(k-4)$;

$C = (k-2)(k-3)(k-4)$; $0 < k < \infty$

2.1.1 Properties of $\phi_j(k)$:

(i) At k = 1; A = 0; B = 0; C = - 6

$$\phi_1'(1) = \bar{y}_r \frac{\bar{x}'}{\bar{x}}; \quad \phi_2'(1) = \bar{y}_r \frac{\bar{x}}{\bar{x}_r}; \quad \phi_3'(1) = \bar{y}_r \frac{\bar{x}'}{\bar{x}_r}$$

(ii) At k = 2; A = 0; B = -2; C = 0

$$\phi_3'(2) = \bar{y}_r \frac{\bar{x}}{\bar{x}'}; \quad \phi_2'(2) = \bar{y}_r \frac{\bar{x}_r}{\bar{x}}; \quad \phi_3'(2) = \bar{y}_r \frac{\bar{x}_r}{\bar{x}'}$$

(iii) At k = 3; A = 2; B = - 2; C = 0

$$\phi_1'(3) = \bar{y}_r \left[\frac{\bar{x}' - f\bar{x}}{(1-f)\bar{x}'}\right]; \phi_2'(3) = \bar{y}_r \left[\frac{\bar{x} - f\bar{x}_r}{(1-f)\bar{x}}\right]; \ \phi_3'(3) = \bar{y}_r \left[\frac{\bar{x}' - f\bar{x}_r}{(1-f)\bar{x}'}\right]$$

(iv) At k = 4; A = 6; B = 0; C = 0

$$\phi_1'(4) = \phi_2'(4) = \phi_3'(4) = \bar{y}_r$$

Theorem 2.1: The point estimate for S of $\bar{Y}$ are:

$$(\bar{y}'_{FTC})_j = k\bar{y}_r + (1-k)\phi_j'(k);\ j = 1,2,3 \qquad \ldots(2.2)$$

Proof: $\left(\bar{y}'_{FTC}\right)_j = \left(\bar{y}_s\right)_j = \frac{1}{n}\sum_{i \in S}(y_{ji})$

$$= \frac{1}{n}\left[\sum_{i \in R}(y_{ji}) + \sum_{i \in R^c}(y_{ji})\right]$$

$$= \frac{1}{n}\left[\sum_{i \in R}\left\{\left(\frac{kn}{r}\right) y_i + (1-k)\phi_j'(k)\right\} + \sum_{i \in R^c}(1-k)\,\phi_j'(k)\right]$$

$$\left(\bar{y}'_{FTC}\right)_j = k\bar{y}_r + (1-k)\phi'_j(k); \qquad j = 1,2,3$$

2.2.0 Some Special Cases:

$At\ k=1,$ $\left(\bar{y}'_{FTC}\right)_j = \bar{y}_r \quad j=1,2,3$ ...(2.3)

$At\ k=2,$ $\left(\bar{y}'_{FTC}\right)_1 = \bar{y}_r\left(2-\frac{\bar{x}}{\bar{x}'}\right)$ ...(2.4)

$$\left(\bar{y}'_{FTC}\right)_2 = \bar{y}_r\left(2-\frac{\bar{x}_r}{\bar{x}}\right)$$ ...(2.5)

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{y}_r\left(2-\frac{\bar{x}_r}{\bar{x}'}\right)$$ ...(2.6)

$At\ \ k=3,$ $\left(\bar{y}'_{FTC}\right)_1 = \bar{y}_r\left(3-\frac{2(\bar{x}'-f\bar{x})}{(1-f)\bar{x}'}\right)$ ...(2.7)

$$\left(\bar{y}'_{FTC}\right)_2 = \bar{y}_r\left(3-\frac{2(\bar{x}-f\bar{x}_r)}{(1-f)\bar{x}}\right)$$ ...(2.8)

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{y}_r\left(3-\frac{2(\bar{x}'-f\bar{x}_r)}{(1-f)\bar{x}'}\right)$$ ...(2.9)

$At\ \ k=4,$ $\left(\bar{y}'_{FTC}\right)_j = \bar{y}_r \qquad j=1,2,3$ ...(2.10)

3. BIAS AND MEAN SQUARED ERROR

Let B(.) and M(.) denote the bias and mean squared error (M.S.E.) of an estimator under a given sampling design. The large sample approximations are

$$\bar{y}_r = \bar{Y}(1+e_1);\ \bar{x}_r = \bar{X}(1+e_1), \bar{x} = \bar{X}(1+e_3);\ \ \bar{x}' = \bar{X}(1+e'_3)$$ ...(3.1)

Using the concept of two phase sampling following Rao and Sitter (1995) and the mechanism of MCAR for given r, n and n'. we have

$$\left.\begin{aligned} &E(e_1) = E(e_2) = E(e_3) = E(e'_3) = 0 \\ &E(e_1^2) = \delta_1 C_Y^2;\ E(e_2^2) = \delta_1 C_X^2;\ E(e_3^2) = \delta_2 C_X^2; E(e_3'^2) = \delta_3 C_X^2; \\ &E(e_1e_2) = \delta_1\rho C_Y C_X; E(e_1e_3) = \delta_2\rho C_Y C_X; E(e_1e'_3) = \delta_3\rho C_X C_Y; \\ &E(e_2e_3) = \delta_2 C_X^2;\ E(e_2e'_3) = \delta_3 C_X^2;\ E(e_3e'_3) = \delta_3 C_X^2 \end{aligned}\right\}$$ ...(3.2)

where $\delta_1 = \left(\frac{1}{r}-\frac{1}{n'}\right);\ \delta_2 = \left(\frac{1}{n}-\frac{1}{n'}\right); \delta_3 = \left(\frac{1}{n'}-\frac{1}{N}\right)$

Theorem 3.1: Estimator $\left(\bar{y}'_{FTC}\right)_j;\ j=1,2,3$ in terms of $e_i, i=1,2,3$ and e'_i could be expressed as:

(i) $\left(\bar{y}'_{FTC}\right)_1 = \bar{Y}\left[1+e_1+(1-k)P\{e_3-e'_3+e_1e_3-e_1e'_3-(\theta_3-\theta_4)e_3e'_3-\theta_4e_3^2+\theta_3e_3'^2\}\right]$ ...(3.3)

(ii) $\left(\bar{y}'_{FTC}\right)_2 = \bar{Y}\left[1+e_1+(1-k)P\{e_2-e_3+e_1e_2-e_1e_3-(\theta_3-\theta_4)e_2e_3-\theta_4e_2^2+\theta_3e_3^2\}\right]$ ...(3.4)

(iii) $\left(\bar{y}'_{FTC}\right)_3 = \bar{Y}\left[1+e_1+(1-k)P\{e_2-e'_3+e_1e_2-e_1e'_3-(\theta_3-\theta_4)e_2e'_3-\theta_4e_2^2+\theta_3e_3'^2\}\right]$ ...(3.5)

Proof :

(i) $\left(\bar{y}'_{FTC}\right)_1 = k\bar{y}_r + (1-k)\phi_1(k)$

Since

$$\phi_1'(k) = \bar{y}_r\left[\frac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}}\right] = \bar{Y}(1+e_1)\left[\frac{(A+fB+C)+(A+C)e_3' + fBe_3}{(A+fB+C)+(A+fB)e_3' + Ce_3}\right]$$

$$= \bar{Y}(1+e_1)\left[\frac{1+\theta_1 e_3' + \theta_2 e_3}{1+\theta_3 e_3' + \theta_4 e_3}\right] = \bar{Y}(1+e_1)(1+\theta_1 e_3' + \theta_2 e_3)(1+\theta_3 e_3' + \theta_4 e_3)^{-1}$$

$$\left[Note:-Binomial\ theorem\ (1+\alpha e)^{-1} = 1-\alpha e + \alpha^2 e^2 - \alpha^3 e^3 + ...\right]$$

$$= \bar{Y}(1+e_1)(1+\theta_1 e_3' + \theta_2 e_3)[1-(\theta_3 e_3' + \theta_4 e_3) + (\theta_3 e_3' + \theta_4 e_3)^2 + ...]$$

$$\phi_1(k) = \bar{Y}\left[1+e_1+P\left\{e_3 - e_3' + e_1e_3 - e_1e_3' - (\theta_3-\theta_4)e_3e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2\right\}\right]$$

Therefore,

$$(\bar{y}'_{FTC})_1 = \bar{Y}\left[1+e_1+(1-k)P\left\{e_3 - e_3' + e_1e_3 - e_1e_3' - (\theta_3-\theta_4)e_3e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2\right\}\right]$$

(ii): $\left(\bar{y}'_{FTC}\right)_2 = k\bar{y}_r + (1-k)\phi_2(k)$

$$\phi_2(k) = \bar{Y}_r\left[\frac{(A+C)\bar{x} + fB\bar{x}_r}{(A+fB)\bar{x} + C\bar{x}_r}\right] = \bar{Y}(1+e_1)\left[\frac{1+\theta_1 e_3 + \theta_2 e_2}{1+\theta_3 e_3 + \theta_4 e_2}\right]$$

$$= \bar{Y}(1+e_1)\left[(1+\theta_1 e_3 + \theta_2 e_2)(1+\theta_3 e_3 + \theta_4 e_2)^{-1}\right]$$

$$= \bar{Y}(1+e_1)\left[1+(\theta_1-\theta_3)e_3 + (\theta_2-\theta_4)e_2 + (\theta_4-\theta_2)\theta_4 e_2^2 - (\theta_2\theta_3 + \theta_1\theta_4 - 2\theta_3\theta_4)e_2e_3 - (\theta_1-\theta_3)\theta_3 e_3^2\right]$$

$$= \bar{Y}\left[1+e_1+P(e_2 - e_3 - \theta_4 e_2^2 + \theta_3 e_3^2 - (\theta_3-\theta_4)e_2e_3 + e_1e_2 - e_1e_3)\right]$$

$$= \bar{Y}\left[1+e1+P(e_2 - e_3 + e_1e_2 - e_1e_3 - (\theta_3-\theta_4)e_2e_3 - \theta_4 e_2^2 + \theta_3 e_3^2)\right]$$

Hence $\left(\bar{y}_{FTC}\right)_2 = \bar{Y}\left[(1+e_1) + (1-k)P(e_2 - e_3 + e_1e_2 - e_1e_3 - (\theta_3-\theta_4)e_2e_3 - \theta_4 e_2^2 + \theta_3 e_3^2)\right]$

(iii) : $\left(\bar{y}_{FTC}\right)_3 = k\bar{y}_r + (1-k)\phi_3(k)$

$$\phi_3(k) = \bar{y}_r\left[\frac{(A+C)\bar{x}' + fB\bar{x}_r}{(A+fB)\bar{x}' + \bar{x}_r}\right] = \bar{Y}(1+e_1)\left[(1+\theta_1 e_3'\theta_2 e_2)(1+\theta_3 e_3' + \theta_4 e_2)\right]$$

$$= \bar{Y}(1+e_1)\left[1 - Pe_3' + Pe_2 + P\theta_3 e_3'^2 - P\theta_4 e_2^2 - P(\theta_3-\theta_4)e_2e_3'\right]$$

$$= \bar{Y}\left[1+P(e_2 - e_3' + \theta_3 e_3'^2 - \theta_4 e_2^2 - (\theta_3-\theta_4)e_2e_3') + e_1 + P(e_1e_2 - e_1e_3' + \theta_3 e_1 e_3'^2 - \theta_4 e_1 e_2^2 - (\theta_3-\theta_4)e_1e_2e_3')\right]$$

$$= \bar{Y}\left[1+e_1+P(e_2 - e_3' + e_1e_2 - e_1e_3' - (\theta_3-\theta_4)e_2e_3' - \theta_4 e_2^2 + \theta_3 e_3'^2)\right]$$

Hence,

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{Y}\left[(1+e_1) + (1-k)P(e_2 - e_3' + e_1e_2 - e_1e_3' - (\theta_3-\theta_4)e_2e_3' - \theta_4 e_2^2 + \theta_3 e_3'^2)\right]$$

Theorem (3.2): The bais of the estimators $\left(\bar{y}'_{FTC}\right)_j$ is given by

(i) $$B\left[\left(\bar{y}'_{FTC}\right)_1\right] = -\bar{Y}P(1-k)(\delta_2-\delta_3)\left[\theta_4 C^2{}_X - \rho C_Y C_X\right]$$

(ii) $$B\left[\left(\bar{y}'_{FTC}\right)_2\right] = -\bar{Y}(1-k)P(\delta_1-\delta_2)\left[\theta_4 C_x^2 - \rho C_Y C_X\right]$$

(iii) $$B\left[\left(\bar{y}'_{FTC}\right)_3\right] = -\bar{Y}(1-k)P(\delta_1-\delta_3)\left[\theta_4 C_X^2 - \rho C_Y C_X\right]$$

Proof:

(i): $$B\left[\left(\bar{y}'_{FTC}\right)_1\right] = E\left[\left(\bar{y}'_{FTC}\right)_1 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{1+e_1+(1-k)P(e_3-e_3'+e_1e_3-e_1e_3'-(\theta_3-\theta_4)e_3e_3'-\theta_4e_3^2+\theta_3e_3'^2)\right\}-\bar{Y}\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_2-\delta_3)\rho C_Y C_X - \{(\theta_3-\theta_4)\delta_3+\theta_4\delta_2\}C_X{}^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_2-\delta_3)\rho C_Y C_X - (\delta_2-\delta_3)\theta_4 C_X^2\right]$$

$$= -\bar{Y}P(1-k)(\delta_2-\delta_3)\left[\theta_4 C^2{}_X - \rho C_Y C_X\right] \quad \ldots(3.6)$$

(ii) $$B\left[\left(\bar{y}'_{FTC}\right)_2\right] = E\left[\left(\bar{y}'_{FTC}\right)_2 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{1+e_1+(1-k)P(e_2-e_3+e_1e_2-e_1e_3-(\theta_3-\theta_4)e_2e_3-\theta_4e_2^2+\theta_3e_3^2)\right\}-\bar{Y}\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - \{(\theta_3-\theta_4)\delta_2+\theta_4\delta_1-\theta_3\delta_2\}C_X^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - \{\theta_3\delta_2-\theta_4\delta_2+\theta_4\delta_1-\theta_3\delta_2\}C_X^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - (\delta_1-\delta_2)\theta_4 C_X^2\right]$$

$$= -\bar{Y}(1-k)P(\delta_1-\delta_2)\left[\theta_4 C_x^2 - \rho C_Y C_X\right] \quad \ldots(3.7)$$

(iii) $$B\left[\left(\bar{y}'_{FTC}\right)_3\right] = E\left[\left(\bar{y}'_{FTC}\right)_3 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{(1+e_1)+(1-k)P(e_2-e_3'+e_1e_2-e_1e_3'-(\theta_3-\theta_4)e_2e_3'-\theta_4e_2^2+\theta_3e_3'^2)\right\}-\bar{Y}\right]$$

$$= \bar{Y}\left[(1-k)P\left[(\delta_1-\delta_3)\rho C_Y C_X - \{(\theta_3-\theta_4)\delta_3+\theta_4\delta_1-\theta_3\delta_3\}C_x^2\right]\right]$$

$$= -\bar{Y}(1-k)P(\delta_1-\delta_3)\left[\theta_4 C_X^2 - \rho C_Y C_X\right] \quad \ldots(3.8)$$

Theorem 3.3: The m.s.e. of the estimators $\left(\bar{y}'_{FTC}\right)_j$ is given by:-

(i) $$M\left[(\bar{y}'_{FTC})_1\right] = \bar{Y}\left[\delta_1 C_Y^2 + (1-k)^2 P^2(\delta_2-\delta_3)C_x^2 + 2(1-k)P(\delta_2-\delta_3)e^{C_Y C_X}\right] \quad \ldots(3.9)$$

(ii) $$M\left[(\bar{y}'_{FTC})_2\right] = \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2 P^2(\delta_1-\delta_2)C_X^2 + 2(1-k)P(\delta_1-\delta_2)\rho C_Y C_X\right] \ldots(3.10)$$

(iii) $$M\left[(\bar{y}'_{FTC})_3\right] = \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2 P^2(\delta_1-\delta_3)C_X^2 + 2(1-k)P(\delta_1-\delta_3)\rho C_Y C_X\right] \ldots(3.11)$$

Proof:

(i): $$M\left[(\bar{y}'_{FTC})_1\right] = E\left[(\bar{y}_{FTC})_1 - \bar{Y}\right]^2$$

Using equation (3.3)

$$= \bar{Y}^2 E\left[e_1+(1-k)P\left\{e_3-e_3'+e_1e_3-e_1e_3'-(\theta_3-\theta_4)e_3e_3'-\theta_4e_3^2+\theta_3e_3'^2\right\}\right]^2$$

$$= \bar{Y}^2 E\left[e_1 + (1-k)P(e_3 - e_3')\right]^2$$
$$= \bar{Y}^2 E\left[e_1^2 + (1-k)^2 P^2 (e_3 - e_3')^2 + 2(1-k)P(e_3 - e_3')e_1\right]$$
$$= \bar{Y}\left[\delta_1 C_Y^2 + (1-k)^2 P^2 (\delta_2 - \delta_3) C_x^2 + 2(1-k)P(\delta_2 - \delta_3)\rho C_Y C_X\right]$$

(ii) $$M\left[(\bar{y}'_{FTC})_2\right] = E\left[(\bar{y}'_{FTC})_2 - \bar{Y}\right]^2$$

From using equation (3.4)

$$= E\left[Y\left\{1 + e_1 + (1-k)P\left\{e_2 - e_3 + e_1 e_2 - e_1 e_3 - (\theta_3 - \theta_4)e_2 e_3 - \theta_4 e_2^2 + \theta_3 e_3^2\right\}\right\} - \bar{Y}\right]^2$$
$$= \bar{Y}^2 E\left[e_1^2 + (1-k)^2 P^2 (e_2 - e_3)^2 + (1-k)P(e_2 - e_3)e_1\right]$$
$$= \bar{Y}^2 E\left[e_1^2 + (1-k)^2 P^2 (e_2^2 + e_3^2 - 2e_2 e_3) + 2(1-k)P(e_1 e_2 - e_1 e_3)\right]$$
$$= \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2 P^2 (\delta_1 - \delta_2) C_X^2 + 2(1-k)P(\delta_1 - \delta_2)\rho C_Y C_X\right]$$

(iii) $$M\left[\left(\bar{y}'_{FTC}\right)_3\right] = E\left[\left(\bar{y}'_{FTC}\right)_3 - \bar{Y}\right]^2$$
$$= \bar{Y}^2 E\left[e_1 + (1-k)P\left\{e_2 - e_3'\right\}\right]^2$$
$$= \bar{Y}^2 E\left[e_1^2 + (1-k)^2 P^2\left\{e_2 - e_3'\right\} + 2(1-k)P(e_2 - e_3')e_1\right]^2$$
$$= \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2 P^2 (\delta_1 - \delta_3) C_X^2 + 2(1-k)P(\delta_1 - \delta_3)\rho C_Y C_X\right]$$

Theorem 3.4: The minimum m.s.e of $\left(\bar{y}'_{FTC}\right)_j$ is

(i) $$M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{mim} = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right]S_Y^2 \quad \ldots(3.13)$$

(ii) $$M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_2)\rho^2\right]S_Y^2 \quad \ldots(3.14)$$

(iii) $$M\left[\left(\bar{Y}'_{FTC}\right)_3\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2 \quad \ldots(3.15)$$

Proof:

(i): $$\frac{d}{d(1-k)P} M\left[(\bar{y}'_{FTC})_1\right] = 0$$

From equation (3.9)

$$\Rightarrow \quad (1-k)PC_X + \rho C_Y = 0 \quad \Rightarrow \quad (1-k)P = -\rho\frac{C_y}{C_x}$$

Therefore from equation (3.9). we have

$$M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{\min} = \bar{Y}^2\left[\delta_1 C_Y^2 - (\delta_2 - \delta_3)\rho^2 C_Y^2\right] \qquad \because C_Y^2 = \left(\frac{S_Y}{\bar{Y}}\right)^2$$

Therefore

$$M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{mim} = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right]S_Y^2$$

(ii) $$\frac{d}{d\left[(1-k)P\right]} M\left[\left(\bar{y}'_{FTC}\right)_2\right] = 0$$

From equation (3.10)

$$\Rightarrow \quad (1-k)PC_x + \rho C_Y = 0 \Rightarrow \quad (1-k)P = -\rho\frac{C_Y}{C_X}$$

Therefore

$$M\left[\left(\bar{y}'_{FTC}\right)^2\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_2)\rho^2\right]S_Y^2$$

(iii) $$\frac{d}{d[(1-k)P]} M\left[\left(\bar{y}'_{FTC}\right)_3\right] = 0$$ From equation (3.11)

$$\Rightarrow \quad (1-k)P = -\rho\frac{C_Y}{C_X} \qquad \text{...(3.16)}$$

Therefore $$M\left[\left(\bar{Y}'_{FTC}\right)_3\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2$$

3.1 Multiple Choices of k :

The optimality condition $P = -V$ provides the equation

$$k^4 - (f - V)k^3 - [(4f + 15) - (f - 8)V]k^2 + [(f - 10) - (5f - 23)V]k + [(4f + 24) + (4f - 22)V] = 0 \qquad \text{...(3.17)}$$

which fourth degree polynomial in terms of k. One can get at most four values of k like k_1, k_2, k_3, k_4 for which m. s. e. is optimal. The best choice criteria is

Step I: Compute $\left|B(T_{FTi})_{k_j}\right|$ for i = 1, 2, 3; j = 1, 2, 3, 4.

Step II: For given i, choose k_j as $\left|B(T_{FTi})_{k_j}\right| = \overset{\min}{j=1,2,3,4}\left[\ \left|B(T_{FTi})_{k_j}\right|\ \right]$

This ultimately gives bias control at the optimal level of m.s.e.

Note 3.1: For given pair of values of (V, f), $0 < V < \infty$; $0 < f < 1$, one can generate a trivariate table of k_1, k_2, k_3, k_4 so as to achieve solution quickly.

Remark 3.2: Reddy (1978) has shown that quantity $V = \rho\frac{C_Y}{C_X}$ is stable over moderate length time period and could be priorly known or guessed by past data. Therefore, pair (f, V) be treated as known and equation (3.13) generates maximum of four roots (some may imaginary) on which optimum level of m.s.e. will be attained.

4. COMPARISON

(i) Let $D_1 = M\left[(\bar{y}'_{FTC})_1\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min} = [\delta_1 - 2\delta_1 + \delta_3]\rho^2\delta_Y^2$

Thus $\left(\bar{y}'_{FTC}\right)_2$ is better than $\left(\bar{y}'_{FTC}\right)_1$ if:

$$D_1 > 0 \Rightarrow [\delta_1 - 2\delta_2 + \delta_3]e^2\delta_Y^2 > 0 \Rightarrow \delta_1 - 2\delta_2 + \delta_3 > 0 \qquad \text{...(4.1)}$$

(ii) Let $D_2 = M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_3\right]_{\min} = [-\delta_2 + \delta_3 + \delta_1 - \delta_3]\rho^2\delta_Y^2$

$$= (\delta_1 - \delta_2)\rho^2\delta_Y^2$$

Thus $\left(\bar{y}'_{FTC}\right)_3$ better than $\left(\bar{y}'_{FTC}\right)_1$ if

$$D_2 > 0 \Rightarrow (\delta_1 - \delta_2)\rho 2\delta_Y^2 > 0 \ \Rightarrow \frac{1}{r} - \frac{1}{n} > 0 \Rightarrow \frac{1}{r} > \frac{1}{n} \Rightarrow n > r \qquad \text{...(4.2)}$$

i.e. the size of sample domain is greater than the size of auxiliary data.

(iii) $D_3 = M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_3\right]_{\min} = [(\delta_2 - \delta_3)\rho^2]\delta_Y^2 \quad = (\delta_2 - \delta_3)\rho^2\delta_Y^2$

Thus $\left(\bar{y}'_{FTC}\right)_3$ is better than $\left(\bar{y}'_{FTC}\right)_2$ if

$D_3 > 0 \Rightarrow (\delta_2 - \delta_3) > 0 \Rightarrow \delta_1 -> \delta_3 \Rightarrow \frac{1}{n} - \frac{1}{n'} > \frac{1}{n'} - \frac{1}{N}$ If $n' \to N$

Then $\frac{1}{n} - \frac{1}{N} > 0 \Rightarrow \frac{1}{n} > \frac{1}{N} \Rightarrow N > n$ …(4.3)

i.e. the size of total data set is greater than the size of sample data set.

5. EMPIRICAL STUDY

The attached appendix A has generated artificial population of size N = 200 containing values of main variable Y and auxiliary variable X. Parameter of this are given below:

$\bar{Y}$ = 42.485; $\bar{X}$ = 18.515; S_Y^2 = 199.0598; S_X^2 = 48.5375; ρ = 0.8652; C_X = 0.3763; C_Y = 0.3321. Using random sample SRSWOR of size n = 50; r = 45; f = 0.25, α = 0.2365. Solving optimum condition $\theta = -V$ [see (3.13)] the equation of power four in k provides only two real values k_1 = 0.8350; k_2 =4.1043. Rest other two roots appear imaginary.

6. SIMULATION

The bias and optimum m.s.e. of proposed estimators under both designs are computed through 50,000 repeated samples n, n' as per design. Computations are in table 6.1.

The simulation procedure has following steps :

Step 1: Draw a random sample S' of size $n' = 110$ from the population of N = 200 by SRSWOR.

Step 2: Draw a random sub-sample of size $n = 50$ from S'.

Step 3: Drop down 5 units randomly from each second sample corresponding to Y.

Step 4: Impute dropped units of Y by proposed methods and available methods and compute the relevant statistic.

Step 5: Repeat the above steps 50,000 times, which provides multiple sample based estimates $\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_{50000}$.

Step 6: Bias of $\hat{y}$ is $B(\hat{y}) = \frac{1}{50000}\sum_{i=1}^{50000}\left[\hat{y}_i - \bar{Y}\right]$

Step 7: M.S.E. of $\hat{y}$ is $M(\hat{y}) = \frac{1}{50000}\sum_{i=1}^{50000}\left[\hat{y}_i - \bar{Y}\right]^2$

Table 6.1 : Comparisons of Estimators

Estimator	***Bias* (.)**	***M*(.)**
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_1}$	0.3313	13.5300
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_2}$	0.0489	3.4729
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_3}$	---	---
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_1}$	0.2686	4.6934

$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_2}$	0.0431	3.2194
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_3}$	---	---
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_1}$	0.5705	14.6633
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_2}$	0.0639	3.5274
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_3}$	---	---

TABLE 1: Bias and Optimum m.s.e. at $k = k_i$ $(i = 1,2)$

7. CONCLUDING REMARKS

The content of this paper has a comparative approach for the three estimators examined under two-phase sampling. The estimator $\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_2}$ is best in terms of mean squared error than other estimators. We can also choose an appropriate value of k for minimum bias from available values of k. Equation (4.1), (4.2) and (4.3) shows the general conditions for showing better performance of any estimator. All suggested methods of imputation are capable enough to obtain the values of missing observations in data warehouse. These methods are useful in the case where two attributes are in quantitative manner and linearly correlate with each other, like, Statistical Database, agricultural database (yield and area under cultivation), banking database (saving and interest),Spatial Databases etc. Therefore, suggested strategies are found very effective in order to replace missing values during the data preprocessing in KDD, so that the quality of the results or patterns mined by data mining methods can be improved.

8. REFERENCES

[1]. U Fayyad, Piatetsky-Shapiro, P.Smyth. ”Knowledge discovery and data mining: Towards a unifying framework”,In Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD), Portland, OR, pp 82–88.1996.

[2]. Piatetsky, Shapiro and J.William, Frawley. “Knowledge discovery in databases”,AAAI Press/MIT Press,1991.

[3]. R.Krishnamurthy, and T.Imielinski. “Research directions in Knowledge Discovery”, SIGMOD Record,20(3):76-78,1991.

[4]. D.Pyle. “Data preparation for data mining”, Morgan Kaufmann Publishers Inc, (1999).

[5]. J. Han, M. Kamber. “Data mining: concepts and techniques”, Academic Press, San Diego, (2001).

[6]. H. P. Kriegel, Karsten, M. Borgwardt, P. Kröge, A. Pryakhin, M. Schubert, A. Zimek, “Future trends in data mining”, Data Min Knowl Disc 15:87–97 DOI 10.1007/s10618-007-0067-9,2007.

[7]. J. Kivinen and H.Mannila. “The power of sampling in knowledge discovery”, In Proc. Thirteenth ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Sys., pages 77–85. ACM Press,1994.

[8]. M. J. Zaki, S. Parthasarathy, W. Lin, and M. Ogihara. “Evaluation of sampling for data mining of association rules”, Technical Report 617, University of Rochester, Rochester, NY,1996.

[9]. H. Toivonen. “Sampling large databases for association rules”, In Proc. 22nd VLDB 1996.

[10]. G. H. John and P. Langley. “Static versus dynamic sampling for data mining”, In Proc. Second Intl. Conf. Knowledge Discovery and Data Mining, pages 367–370. AAAI Press,1996.

[11]. C. Domingo, R. Gavalda and Q. Watanabe. “Adaptive Sampling Methods for Scaling Up Knowledge Discovery Algorithms”, Data Mining and Knowledge Discovery,2002.

[12]. M. Zaki, S. Parthasarathy, W. Li and M. Ogihara. “Evaluation of Sampling for Data Mining of Association Rules”, Proc. Int’l Workshop Research Issues in Data Eng,1997.

[13]. K.T. Chuang, K. P. Lin, and M. S. Chen. “Quality-Aware Sampling and Its Applications in Incremental Data Mining”, IEEE Transactions on knowledge and data engineering,vol.19, no. 4,2007.

[14]. K.Lakshminarayan, S. A. Harp and Samad. “Imputation of missing data in industrial databases, Appl. Intell., vol. 11, no. 3, pp. 259–275, Nov./Dec1999.

[15]. R. J. Little and D. B. Rubin. “Statistical Analysis With Missing Data”, Hoboken, NJ: Wiley, (1987).

[16]. H. L. Oh, and F. L. Scheuren. “Weighting adjustments for unit nonresponse, incomplete data in sample survey”, in Theory and Bibliographies, vol. 2, W. G. Madow, I. Olkin, and D. B. Rubin, Eds. New York: Academic, pp. 143–183,1983.

[17]. W. S. Sarle. “Prediction with missing inputs”, in Proc. 4th JCIS, vol. 2, pp. 399–402,1998.

[18]. K. J. Cios, W. Pedrycz, ,and R. Swiniarski. “Data Mining Methods for Knowledge Discovery”,Norwell, MA: Kluwer,(1998).

[19]. K. Chan, T. W. Lee, and T. J. Sejnowski. “Variational Bayesian learning of ICA with missing data, Neural Comput”, vol. 15, no. 8, pp. 1991–2011,2003.

[20]. Y. Freund and R. E. Schapire. “Experiments with a new boosting algorithm”, in Proc. 13th Int. Conf. Mach. Learn., pp. 146–148,1996.

[21]. V. Tresp, R. Neuneier, and S. Ahmad. “Efficient methods for dealing with missing data in supervised learning”, in Advances in Neural Information Processing Systems 7, G. Cambridge, MA: MIT Press, pp. 689–696,1995.

[22]. W. Zhang. “Association based multiple imputation in multivariate datasets”, A summary, in Proc. 16th ICDE, pp. 310–311,2000.

[23]. J. R. Quinlan. ”C4.5: Programs for Machine Learning”, San Mateo, CA: Morgan Kaufmann,1992.

[24]. J. R. Quinlan. “Induction of decision trees, Mach. Learn”, vol. 1, no. 1, pp. 81–106, 1986.

[25]. A. Farhangfar, L. A. Kurgan, and W. Pedrycz. “Novel framework for imputation of missing values in databases”, Comput.: Theory and Appl. II Conf., Conjunction with SPIE Defense and Security Symp. (formerly AeroSense), Orlando, FL, pp. 172–182,2004.

[26]. G. Batista and M. Monard. “An analysis of four missing data treatment methods for supervised learning”, Appl. Artif. Intell., vol. 17, no. 5/6, pp. 519–533,2003

[27]. W. G. Cochran. “Sampling Techniques”, John Wiley and Sons, New York, (2005).

[28]. D. F. Heitjan and S. Basu. “Distinguishing ‘Missing at random’ and ‘missing completely at random”, The American Statistician, 50, 207-213,1996.

[29]. V. N. Reddy. “A study on the use of prior knowledge on certain population parameters in estimation”, Sankhya, C, 40, 29-37,1978.

[30]. D. Shukla. “F-T estimator under two-phase sampling”, Metron, 59, 1-2, 253-263,2002.

[31]. S. Singh, and S. Horn. “Compromised imputation in survey sampling”, Metrika, 51, 266-276,2000.

[32]. Li.Liu, Y. Tu, Y. Li, and G. Zou. “Imputation for missing data and variance estimation when auxiliary information is incomplete”, Model Assisted Statistics and Applications, 83-94,2005.

[33]. S.Singh. “A new method of imputation in survey sampling”, Statistics, Vol. 43, 5 , 499 – 511,2009.

Appendix A (Artificial Dataset (N = 200))

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19

Applying Statistical Dependency Analysis Techniques In a Data Mining Domain

Sudheep Elayidom *sudheepelayidom@hotmail.com*
Computer Science and Engineering Division
School of Engineering
Cochin University of Science and Technology
Kochi, 682022, India

Sumam Mary Idikkula *umam@cusat.ac.in*
Department of Computer Science
Cochin University of Science and Technology
Kochi, 682022, India

Joseph Alexander *josephalexander@cusat.ac.in*
Project Officer, Nodel Centre
Cochin University of Science and Technology
Kochi, 682022, India

Abstract

Taking wise career decision is so crucial for anybody for sure. In modern days there are excellent decision support tools like data mining tools for the people to make right decisions. This paper is an attempt to help the prospective students to make wise career decisions using technologies like data mining. In India technical manpower analysis is carried out by an organization named NTMIS (National Technical Manpower Information System), established in 1983-84 by India's Ministry of Education & Culture. The NTMIS comprises of a lead center in the IAMR, New Delhi, and 21 nodal centers located at different parts of the country. The Kerala State Nodal Center is located in the Cochin University of Science and Technology. Last 4 years information is obtained from the NODAL Centre of Kerala State (located in CUSAT, Kochi, India), which stores records of all students passing out from various technical colleges in Kerala State, by sending postal questionnaire. Analysis is done based on Entrance Rank, Branch, Gender (M/F), Sector (rural/urban) and Reservation (OBC/SC/ST/GEN).

Key Words: Confusion matrix, Data Mining, Decision tree, Neural Network, Chi- square

1. INTRODUCTION

The popularity of subjects in science and engineering in colleges around the world is up to a large extent dependent on the viability of securing a job in the corresponding field of study. Appropriation of funding of students from various sections of society is a major decision making hurdle particularly in the developing countries. An educational institution contains a large number of student records. This data is a wealth of information, but is too large for any one person to understand in its entirety. Finding patterns and characteristics in this data is an essential task in education research, and is part of a larger task of developing programs that increase student learning. This type of data is presented to decision makers in the state government in the form of tables or

charts, and without any substantive analysis, most analysis of the data is done according to individual intuition, or is interpreted based on prior research. this paper analyzed the trends of placements in the colleges, keeping in account of details like rank, sex, category and location using decision tree models like naive bayes classifier, neural networks etc. chi square test is often shorthand for pearson chi square test.it is a statistical hypothesis test. spss (statistical package for social science) is most widely used programed for statistical analysis. the data preprocessing for this problem has been described in detail in articles [1] & [9], which are papers published by the same authors. The problem of placement chance prediction may be implemented using decision trees. [4] Surveys a work on decision tree construction, attempting to identify the important issues involved, directions which the work has taken and the current state of the art. Studies have been conducted in similar area such as understanding student data as in [2]. there they apply and evaluate a decision tree algorithm to university records, producing graphs that are useful both for predicting graduation, and finding factors that lead to graduation. It's always been an active debate over which engineering branch is in demand .so this work gives a scientific solution to answer these. Article [3] provides an overview of this emerging field clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases.

[5] Suggests methods to classify objects or predict outcomes by selecting from a large number of variables, the most important ones in determining the outcome variable. the method in [6] is used for performance evaluation of the system using confusion matrix which contains information about actual and predicted classifications done by a classification system. [7] & [8] suggest further improvements in obtaining the various measures of evaluation of the classification model. [10] Suggests a data mining approach in students results data while [11] and [12] represents association rules techniques in the data mining domain.

2. DATA

The data used in this project is the data supplied by National Technical Manpower Information System (NTMIS) via Nodal center. Data is compiled by them from feedback by graduates, post graduates, diploma holders in engineering from various engineering colleges and polytechnics located within the state during the year 2000-2003. This survey of technical manpower information was originally done by the Board of Apprenticeship Training (BOAT) for various individual establishments. A prediction model is prepared from data during the year 2000-2002 and tested with data from the year 2003.

3. PROBLEM STATEMENT

To prepare data mining models and predict placement chances for students keeping account of input details like his/her Rank, Gender, Branch, Category, Reservation and Sector. Statistical dependency analysis techniques are to be used for input attributes to determine the attributes on which placement chances are dependent on. Also performances of the models are to be compared.

4. CONCEPTS USED

4.1. Data Mining

Data mining is the principle of searching through large amounts of data and picking out interesting patterns. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases".

A typical example for a data mining scenario may be “In a mining analysis if it is observed that people who buy butter tend to buy bread too then for better business results the seller can place butter and bread together.”

4.2 Cross Tabulation

Count		*Chance*				*Total*
		1	2	3	4	
Sector	1	324	194	68	168	754
	2	416	114	69	192	791
Total		740	308	137	360	1545

TABLE 1: A Sample Cross tabulation

The row and column variables are independent, or unrelated. If that assumption was true one would expect that the values in the cells of the table are balanced. To determine what is meant by *balanced*, consider a simple example with two variables, sector and chance for example. It is to be decided on whether there is a relation between sectors (male/female) and chance (yes/no), or whether the two variables are independent of each other. An experiment is to be conducted by constructing a crosstab table as in table 2.

	chance-1	*chance-2*	*Totals*
sector-1	22	18	*40*
sector-2	26	34	*60*
Totals	*48*	*52*	*100*

TABLE 2: Cross tabulation table 2

Now the sector-1 and sector-2 are divided pretty much evenly among chance-1and chance-2 suggesting perhaps that the two variables *are* independent of each other. Suppose it is decided again to conduct the experiment and select some random sample, but, if only totals for each variable separately are considered, for example:

Number of sector-1 is 30; number of sector-2 is 70

Number of chance-1 is 40; number of chance-2 is 60

Total number of data values (subjects) is 100

With this information we could construct a crosstabs tables as in table 3.

	chance-1	*chance-2*	*Totals*
sector-1			*30*
sector-2			*70*
Totals	*40*	*60*	*100*

TABLE 3: Cross tabulation table 3

Now it is to be understood what kind of distribution in the various cells one would *expect* if the two variables were independent. It is known that 30 of 100 (30%) are sector-1 there are 40 of chance-1and 60 of chance-2- if chance had nothing to do with sector (the variables were independent) that we would expect that 30% of the 40 of chance-1are of sector-1, while 30% of the 60 chance-2 are of sector-1. Same concept may be applied to sector 2 also.

Under the assumption of independence one can *expect* the table to look as in table 4:

	chance-1	*chance-2*	*Totals*
sector-1	30/100 * 40 = 12	30/100 * 60 = 18	*30*
sector-2	70/100 * 40 = 28	70/100 * 60 = 42	*70*
Totals	*40*	*60*	*100*

TABLE 4: Cross tabulation table 4

In other words, if a crosstabs table with 2 rows and 2 columns has a row totals r1 and r2, respectively, and column totals c1 and c2, and then if the two variables were indeed independent one would expect the complete table to look as follows:

	X	*Y*	*Totals*
A	$r_1 * c_1$ / total	$r_1 * c_2$ / total	r_1
B	$r_2 * c_1$ / total	$r_2 * c_2$ / total	r_2
Totals	c_1	c_2	*total*

TABLE 5: Cross tabulation table 5

The procedure to test whether two variables are independent is as follows:

Create a crosstabs table as usual, called the actual or observed values (not percentages)

Create a second crosstabs table where you leave the row and column totals, but erase the number in the individual cells.

If the two variables were independent, the entry in i-th row and j-th column is expected to be,

(TotalOfRow i) * (totalOfColumn j) / (overallTotal)

Fill in all cells in this way and call the resulting crosstabs table the expected values table

The important point to be noted is that, if the actual values are *very different* from the expected values, the conclusion is that the variables can not be independent after all (because if they were independent the actual values should look similar to the expected values).

The only question left to answer is "how different is very different", in other words when it can be decided that actual and expected values are sufficiently different to conclude that the variables are not independent? The answer to this question is the Chi-Square Test.

4.3 The Chi-Square Test

The Chi-Square test computes the sum of the differences between actual and expected values (or to be precise the sum of the squares of the differences) and assign a probability value to that number depending on the size of the difference and the number of rows and columns of the crosstabs table.

If the probability value *p* computed by the Chi-Square test is very small, differences between actual and expected values are judged to be significant (large) and therefore you conclude that the assumption of independence is invalid and *there must be a relation* between the variables. The error you commit by rejecting the independence assumption is given by this value of p.

If the probability value p computed by the Chi-Square test is large, differences between actual and expected values are not significant (small) and you do not reject the assumption of independence, i.e. it is likely that the variables are indeed independent.

	Value	Df	Asymp. Sig (2-sided)
Pearson chi-square	32.957[a]	3	.000
Likelyhood ratio	33.209	3	.000
Linear by linear association	.909	1	.340
N of valid cases	1545		

TABLE 6: Sample Chi Square output from SPSS

4.4 SPSS

SPSS is among the most widely used software for statistical analysis in social science problems. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The original SPSS manual (Nie, Bent & Hull, 1970) has been described as one of "sociology's most influential books".In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the datafile) are features of the base software. SPSS can read and write data from ASCII text files (including hierarchical files), other statistics packages, spreadsheets and databases. SPSS can read and write to external relational database tables via ODBC and SQL. In this project data was fed as MS Excel spread Sheets of data.

4.5 Results of the Chi-Square Test

An "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each observation is allocated to one cell of a two-dimensional array of cells (called a table) according to the values of the two outcomes. If there are r rows and c columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N},$$

And fitting the model of "independence" reduces the number of degrees of freedom by p = r + c − 1. The value of the test-statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

The number of degrees of freedom is equal to the number of cells rc, minus the reduction in degrees of freedom, p, which reduces to (r − 1) (c − 1). For the test of independence, a chi-square probability of less than or equal to 0.05 (or the chi-square statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.

In our analysis, pairs of attributes namely reservation, sex, sector and rank verus the placement chance were having pearson's chi square value less than 0.05. So It could be concluded that this pair of attributes is dependent. In other words placement chances are showing dependency on reservation, sector, sex and rank.

5. DATA PRE PROCESSING

The individual database files(DBF format) for the years 2000-2003 were obtained and one containing records of students from the year 2000-2002 and another for year 2003, were created.

List of attributes extracted:

RANK: Rank secured by candidate in the engineering entrance exam.

CATEGORY: Social background.

Range: {General, Scheduled Cast, Scheduled Tribe, Other Backward Class}

SEX : Range {Male, Female}

SECTOR: Range {Urban, Rural}

BRANCH: Range {A-J}

PLACEMENT: Indicator of whether the candidate is placed.

The data mining models were built using data from years 2000-2002 and tested using data of year 2003.

6. IMPLEMENTATION LOGIC

6.1. Data Preparation

The implementation begins by extracting the attributes RANK, SEX, CATEGORY, SECTOR, and BRANCH from the master database for the year 2000-2003 at the NODAL Centre. The database was not intended to be used for any purpose other maintaining records of students. Hence there

were several inconsistencies in the database structure. By effective pruning the database was cleaned.

A new table is created which reduces individual ranks to classes and makes the number of cases limited. All queries will belong to a fix set of known cases like:

RANK (A) SECTOR (U) SEX (M)

CATEGORY (GEN) BRANCH (A)

With this knowledge we calculate the chance for case by calculating probability of placement for a test case:

Probability (P) = Number Placed/ Total Number

The chance is obtained by the following rules:

If P>=95 Chance='E'

If P>=85 && P<95 Chance='G'

If P>=50 && P<75 Chance='A';

Else Chance='P';

Where E, G, A, P stand for Excellent, Good, Average & Poor respectively.

The important point is that, for each of the different combination of input attribute values, we compute the placement chances as shown in the above conditions and prepare another table, which is the input for building the data mining models.

7 DATA MINING PROCESSES APPLIED TO THE PROBLEM

7.1 Using Naive Bayes Classifier

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

The classifier is based on Bayes theorem, which is stated as:

P (A|B) = P (B|A)*P (A)/P (B)

Each term in Bayes' theorem has a conventional name:

*P (*A*) is the prior probability or marginal probability of *A*. It is "prior" in the sense that it does not take into account any information about *B*.

*P (A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

*P (B|A) is the conditional probability of B given A.

*P (B) is the prior or marginal probability of B, and acts as a normalizing constant.

Bayes' theorem in this form gives a mathematical representation of how the conditional probabability of event A given B is related to the converse conditional probabablity of B given A.

Confusion Matrix					
		P R E D I C T E D			
		E	P	A	G
AC-TUAL	E	496	10	13	0
	P	60	97	12	1
	A	30	18	248	0
	G	34	19	22	3

TABLE 7: Confusion Matrix (student data)

For training, we have used records 2000-2002 and for testing we used the records of year 2003. We compared the predictions of the model for typical inputs from the training set and that with records in test set, whose actual data are already available for test comparisons.

The results of the test are modelled as a confusion matrix as shown in the above diagram, as its this matrix that is usually used to descrbe test results in data mining type of research works.

The confusion matrix obtained for the test data was as follows:

ACCURACY = 844/1063 = 0.7939

To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. In this case we got an accuracy of **83.0%**. The modified Confusion matrix obtained is as follows:

		Predicted	
		Negative	Positive
Actual	Negative	365	101
	Positive	57	540

TABLE 8: Modified Confusion Matrix (student data)

TP = 0.90, *FP* = 0.22, *TN* = 0.78, *FN* = 0.09.

We have used WEKA, the data mining package for testing using Naive Bayes classifier.

7.2 Decision Tree

A decision tree is a popular classification method that results in a tree like structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the classes. Decision tree is a model that is both predictive and descriptive. In data mining and machine learning, a decision tree is a predictive model. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). The machine learning technique for inducing a decision tree from data is called decision tree learning. For simulation/evaluation we use the data of year 2003 obtained from NTMIS. The knowledge discovered is expressed in the form of confusion matrix.

Confusion Matrix					
		P R E D I C T E D			
		P	A	G	E
AC-TUAL	P	30	1	3	86
	A	7	404	4	11
	G	2	1	4	7
	E	74	6	7	416

TABLE 9: Confusion Matrix (student data)

Since the negative cases here are when the prediction was Poor /average and the corresponding observed values were Excellent/good and vice versa. Therefore the accuracy was given by *AC* = 854/1063 = 0.803. To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. Then the observed accuracy was **82.4%**.

TP = 0.84, *FP* = 0.19, *TN* = 0.81, *FN* = 0.16. We implemented and tested the decision tree concept in a web site using php, mysql platform by using data structure called adjascency list to implement a decision tree.

7.3 Neural Network

Neural Network has the ability to realize pattern recognition and derive meaning from complicated or imprecise data that are too complex to be noticed by either humans or other computer techniques. For simulation/evaluation we use the data of year 2003 obtained from NTMIS. The knowledge discovered is expressed in the form of confusion matrix

Confusion Matrix					
		PREDICTED			
		P	A	G	E
ACTUAL	P	31	4	1	91
	A	5	410	2	9
	G	1	1	6	12
	E	72	13	4	401

TABLE 10: Confusion Matrix (student data)

Since the negative cases here are when the prediction was Poor /average and the corresponding observed values were Excellent/good and vice versa.

Therefore the accuracy is given by

AC = 848/1063 = 0.797

To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. Then the observed accuracy was **82.1** %.

TP = 0.83, FP = 0.17, TN = 0.81, FN = 0.17.

We used MATLAB to implement and test the neural network concept.

8. CONCLUSION

Choosing the right career is so important for any one's success. For that, we may have to do lot of history data analysis, experience based assessments etc. Nowadays technologies like data mining is there which uses concepts like naïve Bayes prediction to make logical decisions. This paper demonstrates how Chi Square based test can be used to evaluate attributes dependencies. Hence this work is an attempt to demonstrate how technology can be used to take wise decisions for a prospective career. The methodology has been verified for its correctness and may be extended to cover any type of careers other than engineering branches. This methodology can very efficeiently be implemented by the governments to help the students make career decisions. It was observed that the performances of all the three models were comparable in the domain of placement chance prediction as a part of the original research work.

9. ACKNOWLEDGEMENT

I would like to acknowledge the technical contributions of Sunny(sunny@hotmail.co.in), Vikash kumar(vikashhotice2006@gmail.com),Vinesh.B(vineshbalan@gmail.com),Amit.N(amitnanda@hotmail.com), Vikash Agarwal (vikash.vicky007@gmail.com), Division of Computer Engineering, Cochin University Of Science and Technology, India.

10. REFERENCES

[1] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander. “Applying datamining using statistical techniques for career selection”. IJRTE, 1(1):446-449, 2009

[2] Elizabeth Murray. “Using Decision Trees to Understand Student Data”. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005

[3] Fayyad, R. Uthurusamy. “From Data mining to knowledge discovery”, Advances *in data mining and knowledge discovery* Cambridge, MA: MIT Press., pp. 1-34, (1996)

[4] Sreerama K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, pp. 345-389, 1998

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. “Classification and Regression Trees”, Wadsworth Inc., Chapter 3, (1984.)

[6] Kohavi R. and F. Provost, Editorial for the Special Issue on application of machine learning and the knowledge of discovery process, Machine Learning, 30: 271-274, 1998

[7] M. Kubat, S. Matwin. “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection”. In Proceedings of the 14th International Conference on Machine Learning, ICML, Nashville, Tennessee, USA, 1997

[8] Lewis D. D. & Gale W. A. “A sequential algorithm for training text classifiers”. In proceedings of SIGIR, Dublin, Ireland, 1994

[9] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander. “Applying Data mining techniques for placement chance prediction”. In Proceedings of international conference on advances in computing, control and telecommunication technologies, India, 2009

[10] Oyelade, Oladipupo, Olufunke. “Knowledge Discovery from students result repository: Association Rules mining approach”. CSC- IJCSS, 4(2):199-207, 2010

[11] Anandavalli, Ghose, Gauthaman. “Mining spatial gene expression data using association rules”. CSC - IJCSS, 3(5): 351-357, 2009

[12] Anyanwn, Shiva. “Comparitive analysis of serial decision tree classification algorithms”. CSC- IJCSS, 3(3):230-240, 2009

Modeling of Nitrogen Oxide Emissions in Fluidized Bed Combustion Using Artificial Neural Networks

Mika Liukkonen mika.liukkonen@uef.fi
Department of Environmental Science
University of Eastern Finland
P.O. Box 1627, FIN-70211 Kuopio, Finland

Eero Hälikkä
Foster Wheeler Power Group
P.O. Box 201, FIN-78201 Varkaus, Finland

Reijo Kuivalainen
Foster Wheeler Power Group
P.O. Box 201, FIN-78201 Varkaus, Finland

Yrjö Hiltunen yrjo.hiltunen@uef.fi
Department of Environmental Science
University of Eastern Finland
P.O. Box 1627, FIN-70211 Kuopio, Finland

Abstract

The reduction of harmful emissions is affecting increasingly the modern-day production of energy, while higher objectives are set also for the efficiency of combustion processes. Therefore it is necessary to develop such data analysis and modeling methods that can respond to these demands. This paper presents an overview of how the formation of nitrogen oxides (NO_x) in a circulating fluidized bed (CFB) boiler was modeled by using a sub-model -based artificial neural network (ANN) approach. In this approach, the process data is processed first by using a self-organizing map (SOM) and k-means clustering to generate subsets representing the separate process states in the boiler. These primary process states represent the higher level process conditions in the combustion, and can include for example start-ups, shutdowns, and idle times in addition to the normal process flow. However, the primary states of process may contain secondary states that represent more subtle phenomena in the process, which are more difficult to observe. The data from secondary combustion conditions can involve information on e.g. instabilities in the process. In this study, the aims were to identify these secondary process states and to show that in some cases the simulation accuracy can be improved by creating secondary sub-models. The results show that the approach presented can be a fruitful way to get new information from combustion processes.

Keywords: Fluidized bed, Artificial neural network, Self-organizing map, Multilayer perceptron, Clustering

1. INTRODUCTION

Nowadays the efficiency of energy plants is considered an important topic due to environmental issues and increasing production costs. Efficient combustion of fuels with lower emissions is a challenging task in the modern-day production of energy, especially when inhomogeneous fuels such as coal, bark, or biomass are used. Fortunately, process data can involve important information on the behavior of the process and on different phenomena that affect the emissions and the energy efficiency of a combustion process. This information is valuable when optimizing the process. For this reason, new methods are needed for data processing, analysis and modeling.

Artificial neural networks (ANN) have shown their usability and power in the modeling of industrial processes [1]–[4]. ANNs have provided serviceable applications in diverse fields of industry, for example in energy production, chemical and electronics industry, and even in waste water treatment [5]–[11]. The variety of neural network applications is wide due to their strong advantages including flexibility, nonlinearity, adaptivity, applicability, a high computing power and a high tolerance of faults [2], [12], [13]. These benefits make ANNs a valuable choice for modeling method in industrial processes.

The use of a self-organizing map (SOM) [12] in the analysis of process states has produced a variety of applications in the past years. In 1992, Kasslin et al. [14] have first introduced the concept of process states by using a SOM to monitor the state of a power transformer. Later on, Alhoniemi et al. [15] have broadened the field of SOM-based applications by using the method in the monitoring and modeling of several industrial processes.

Furthermore, our earlier findings [8], [16] support the fact that different states of the fluidized bed combustion process can be discovered in process data. In reality these states can be for instance start-ups, shut-downs, idle times and different states of normal process flow. The behavior of the process, e.g. the quantities of different emission components, can be diverse between these conditions. However, these upper level process states also include secondary process states, where for example the bed temperature is unstable or the steam flow is lower than usual. It is momentous to learn to identify these states as well, because the performance of the process can fluctuate also in a smaller scale but regardless in a way that has an effect on the model accuracy.

Studying sub-models in parallel to more generic models opens an interesting view to modern-day process analysis. This is because process states and their corresponding sub-models can include valuable information on the performance of the process, as our earlier results concerning the activated sludge treatment and the wave soldering process indicate [6], [11]. The sub-model - based approach is a realistic option for instance in cases where it seems apparent that less detectable but still important phenomena are hidden under the generic behavior of the data. In spite of being more difficult to recognize, these phenomena can have a substantial effect on certain events in the combustion process. Fortunately, this kind of underlying information can be exposed by identifying different process states and creating sub-models, progressing from universal to more detailed models.

In this study, ANNs were used to identify the different states of a circulating fluidized bed (CFB) process and to create sub-models where the nitrogen oxide content of the flue gas was simulated. The obtained sub-models were then compared to the generic process model to see whether the accuracy of simulation could be improved by using this method. In the approach used, self-organizing maps [12], k-means clustering [17] and multilayer perceptrons [2] were combined sequentially to form an ANN method that benefits the good characteristics of all these methods.

2. PROCESS AND DATA

Fluidized bed combustion is a widely-used combustion technology used in power plants and designed primarily for solid fuels. A typical circulating fluidized bed (CFB) boiler consists of a combustion chamber, a separator and a return leg for the recirculation of the bed particles. The fluidized bed is characteristically composed of sand, fuel ash and a matter for capturing the sulfur. This mixture is fluidized by the primary combustion air brought in from the bottom of the chamber. Because of high fluidizing velocities, the bed particles are persistently moving with the flue gases. The particles are driven through the main combustion chamber into a separator, where the larger particles are extracted and returned to the combustion chamber. Meanwhile, the finer particles are separated from the circulation and removed from the flue gases by a bag-house filter or an electrostatic precipitator located downstream from the boiler's convection section.

One of the advantages of fluidized bed combustion is the large heat capacity of the bed, which ensures steady combustion. Only the start-ups involve the use of supporting fuels such as oil or gas. The purpose of the strong turbulence in the circulating fluidized bed is to support the mixing and combustion of fuel. The typical combustion temperature in CFB boilers is between 850 and 900 °C. The process data from the coal-burning CFB under study comprised 10 000 data rows with a 15 minute time interval, the number of variables being 36.

3. METHODS

3.1 Self-Organizing maps (SOM)

Kohonen's self-organizing map (SOM) [12] is a well-known unsupervised artificial neural network algorithm. The common purpose of SOM is to facilitate data analysis by mapping *n*-dimensional input vectors to structural units called *neurons* for example in a two-dimensional lattice (map). The map reflects variations in the statistics of the data set and selects common features which approximate to the distribution of the data samples. On the SOM, the input vectors with common features are associated with the same or neighboring neurons, which preserves the topological organization of the input data. The common properties of a map neuron can be presented with an n-dimensional, neuron-specific reference vector (prototype vector). The size of the map, or the number of neurons, can be altered depending on the purpose; the more neurons, the more details appear.

The SOM analysis is premised on unsupervised learning. At first, the preliminary reference vectors are initialized randomly by sampling their values from an even distribution whose limits are defined by the input data. During learning the input vectors are then grouped one by one into best matching units (BMU) on the map. The BMU is the neuron whose reference vector has the smallest *n*-dimensional Euclidean distance to the input vector. At the same time, the nearest neighbors of the activated neuron become likewise activated according to a predefined neighborhood function (e.g. Gaussian distribution) that is dependent on the network topology. At the final phase, the reference vectors of all activated neurons are updated.

In this study, the SOM was used as a pre-processor to compress information, to remove noise, and to visualize the data. The fluidized bed boiler data were coded into inputs for a self-organizing network, and a SOM having 384 neurons in a 24 x 16 hexagonal arrangement was constructed. The linear initialization and batch training algorithm were used in training, and the neighborhood function was Gaussian. The map was taught with 10 epochs, and the initial neighborhood had the value of 6. The SOM Toolbox (http://www.cis.hut.fi/projects/somtoolbox/) was used in the analysis under a Matlab version 7.6 software (Mathworks Inc., Natick, MA, USA, 2008) platform.

3.2 Identification of Process States

K-means clustering [17] was used to cluster the SOM reference vectors. The algorithm is initialized by randomly defining *k* cluster centers, and then directing each data point to the cluster

whose mean value is closest to it. The Euclidean distance is generally used as a distance measure in these comparisons. At the next stage, the mean vectors of the data points assimilated to each cluster are calculated and used as new centers in an iterative process. Clustering was performed twice, first to reveal the primary process states, and secondly to discover the secondary process states within the primary states.

There are several computational methods to determine the optimal number of clusters, one of the mostly used being the Davies-Bouldin -index [18]. Small values of the DB-index refer to compact clusters whose centers are far from each other. Hence the optimal number of clusters is the number where the index reaches its minimum. The computation of the optimal cluster structure is useful because thus it is not necessary to know the clusters beforehand.

The difference between two separate cluster center vectors represents the factors that separate the clusters. Therefore, this operation can be used to identify the reasons for the determination of different process states. In practice, this calculation is performed so that the comparable individual vector components of the center vectors are subtracted from each other. This produces a vector of differences, which can be used to identify the clusters as process states.

3.3 Multilayer perceptrons (MLP)

Multilayer perceptrons (MLP) are a widely-used [2], [3], [19], [20], supervised ANN methodology that can be used for example to create prediction models. MLPs consist of processing elements and connections. The processing elements are comprised of at least three layers: an input layer, one or more hidden layers, and an output layer. In training, the inputs are processed forward through consecutive layers of neurons. At first, the input layer distributes the input signals to the first hidden layer. Next, the neurons in the hidden layer process the inputs based on given weights, which can either weaken or strengthen the effect of each input. The weights are determined by a supervised learning process, which means learning from examples, or data samples. The inputs are next processed by a transfer function, which is usually nonlinear, providing the method with the ability to distinguish data that are not linearly separable. After that, the hidden neurons transfer the outcome as a linear combination to the next layer, which is generally the output layer. At the last stage, the performance of the model is evaluated with an independent validation data set. This is done by simulating the known data samples by using the model, and comparing the simulated output values with the real ones by using some model performance measure.

MLP neural networks must be trained to a given problem. One of the mostly used training techniques is the back-propagation [21] algorithm. In back-propagation, the network's output value is compared with the original sample to compute the value of a predefined error function. In the iterative training process the network weights are adjusted to a level that minimizes the error value between the actual and expected output for all input patterns.

A MLP network consisting of an input layer, an output layer and a hidden layer with 15 hidden neurons was used to simulate the flue gas nitrogen oxide (NO_x) content in the primary and secondary process states defined earlier by clustering. Variance scaling was used for pre-processing the data. The pre-processed data was divided into training, training test and validation sets. The training data set, being 60 % of the total 10 000 samples was used for training the network, while 20 % of the data set was put to the separate test set to be used in the back-propagation error calculations. The validation data set included the remaining 20 % of the samples, and was used as an independent test set for testing the model.

The parameters for the neural network were determined experimentally. The hyperbolic tangent sigmoid (*tansig*) transfer function was used in the hidden layer, and the linear (*purelin*) transfer function in the output layer. The resilient back-propagation (*trainrp*) algorithm was operated in the training procedure, and the mean squared error (*mse*) as the error function in training. Matlab version 7.6 (Mathworks Inc., Natick, MA, USA) software with the Neural Network Toolbox (version 6.0) was used in data processing. At the final stage, the prediction performances of the primary

and secondary models were compared to prove the usefulness of the presented method. Index of agreement [22] was used as the model performance measure.

4. RESULTS

The identification of the primary and two examples on the identification of the secondary process states in the CFB process are illustrated in Figure 1. Figure 2 and Table 1 present the NO_x prediction performance of the main model and the sub-models of the levels 1 and 2 in clusters III and 7.

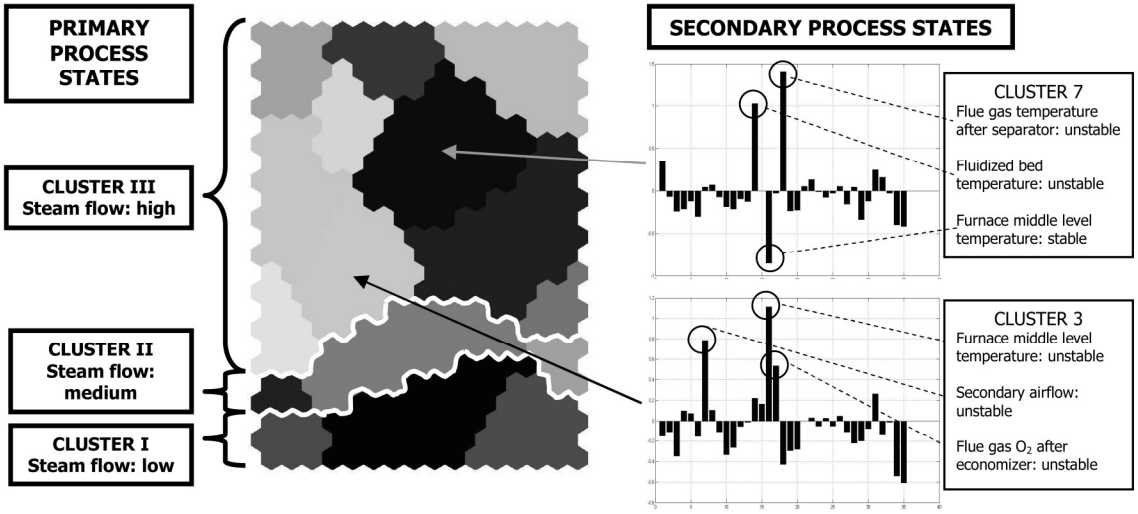

FIGURE 1: Example on the identification of primary and secondary process states on the SOM. The bar graphs represent the difference between the cluster center vector of the secondary process state and the cluster center vector of the primary process state (high steam flow). A considerable positive difference of standard deviations indicates instability in the corresponding variable.

Main model	Level 1 sub-model (Cluster III)	Level 2 sub-model (Cluster 7)
0.950	0.956	0.979

TABLE 1: The goodness (index of agreement) of models a, b and c presented in Figure 2.

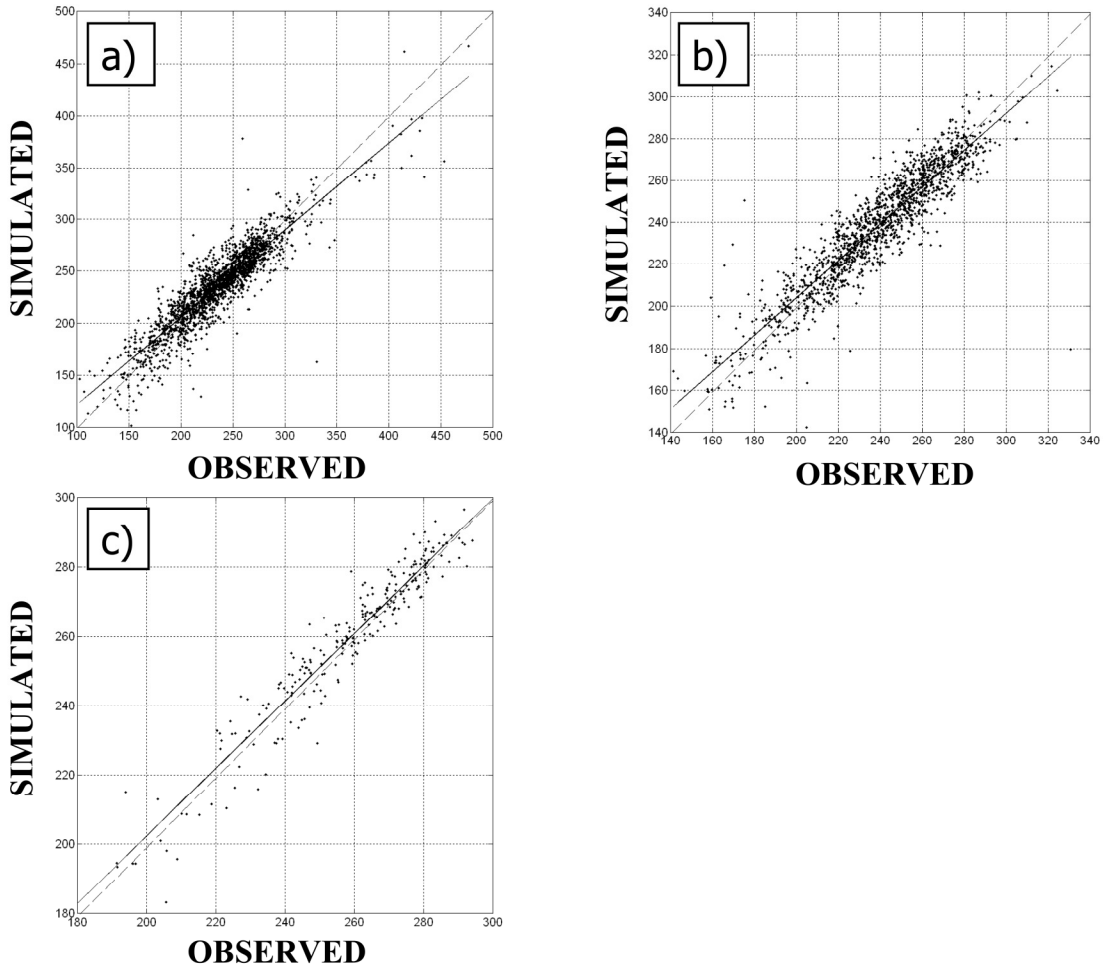

FIGURE 2: Simulated vs. observed flue gas NO_x content in different models. a) indicates the generic main model where the whole data is involved, b) is the level 1 sub-model (Cluster III, high steam flow), and c) is the level 2 sub-model (Cluster 7). Solid line describes least squares and dotted line perfect fitting.

5. DISCUSSION

Present-day energy industry is confronting many challenges, including the tightening legislation on reducing environmental pollution, pressure to increase the energy efficiency of energy plants, and new fuels that are demanding in terms of efficient combustion. The situation is complicated because changes in the combustion process may cause phenomena that have surprising side-effects. For instance, the process may fluctuate to an unstable state where the combustion is inefficient. Unfortunately, an inefficient conversion of fuel to energy often leads to an increased level of emissions. On the other hand, there are also situations where the process works optimally regarding to the efficiency and stability of combustion. It is useful for energy plants if the archived process data can be exploited for advanced process monitoring and diagnostics.

The SOM has been considered a functional, visual and efficient method when it comes to process monitoring, diagnostics, and optimization of circulating fluidized beds [5], [8]–[9], [16]. The findings presented in this study support our earlier results. As a whole, the SOM and k-means provide a simple and functional means to visualize and study the combustion process. Moreover, the results show that the method can be used to define process states and illustrate them in a visual way. All together, it is practical to observe the alteration of process states, because they

provide supplementary information on the operational use of the energy plant and its side-effects, e.g. changes in the levels of nitrogen oxide emission.

The behavior of the CFB process data showed strong distribution of the process events into three principal process states. These states are defined best by high, low and medium steam flow, as can be seen in Figure 1. This kind of principal clustering behavior can be referred to as separation into primary process states. However, the strong grouping of data samples caused by the characteristics of the data means that some interesting phenomena can remain undetected under those strongly defined states. Despite their inconspicuousness, these phenomena can be important in respect to the performance of the combustion process and certain combustion-related events like the formation of emissions. In Figure 1, we have presented one way of revealing this sort of underlying information by identifying the secondary process states. The identification is possible by comparing the cluster center vectors as shown in the figure.

In the literature, experimental models have been successfully developed to model the formation of NO_x in fluidized beds [23]–[26]. In this study we managed to gain good generic models (IA = 0.95) by exploiting real operational data from a coal-fired CFB boiler. However, the accuracy of the NO_x simulations can be improved even as much as 3 % by creating secondary states of process and their sub-models, as can be seen in Figure 2 and Table 1. This means that the model can be improved gradually by getting deeper into the process states and their corresponding sub-models. In this respect, the results support our results concerning the wave soldering and the activated sludge treatment processes [6], [11].

The results suggest that it is reasonable to perform clustering in several successive steps to reveal the unnoticeable phenomena that otherwise may not be recognized. These hidden phenomena are not automatically perceivable without creating sub-models, but may still have a significant effect on certain factors in the process. In other words, the model of a more general level can be sufficient in certain situations, but a more detailed model can be more suitable for describing other phenomena.

In summary, the data analysis scheme used is illustrated in Figure 3. Firstly, the raw process data are pre-processed. Preprocessing includes all the necessary actions, such as data filtering and normalization of variables, to prepare the data for modeling. Next, the SOM and k-means clustering are used and combined with expert process knowledge to obtain the primary process states and their MLP sub-models. After this, the secondary states of process and their corresponding sub-models are formed using the same approach.

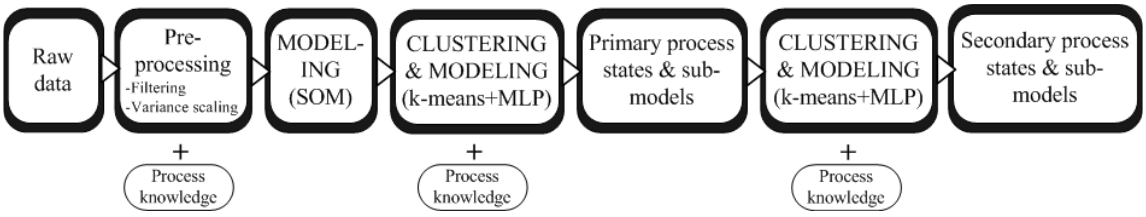

FIGURE 3: The data processing chain in a nutshell

The results show that the universal-to-detailed data analysis method presented can be excellent in cases where high simulation accuracies are required. In addition, the gradual penetrating analysis ensures that the best model is always found for each case in spite of its specific level. The classification of data samples into different subcategories, or process states, provides extra accuracy to the emission model. Furthermore, the ability of the method to reveal nonlinear and multivariate interactions entails additional value to the model. For these reasons, the data analysis method presented offers a powerful option to model industrial processes.

6. CONCLUSIONS

It is apparent that in the future also the energy plants have to be capable of producing energy with a lesser amount of harmful process emissions. Developing new data-based modeling methods for the energy industry is important because there is a growing need for improving the energy conversion processes. In this sense, the utilization possibilities of the approach presented include a wide spectrum of applications. One of them is process diagnostics, which exploits measurement data and has become an essential part of process improvement. Alternatively, the method provides an option for precise modeling of emissions in circulating fluidized bed boilers. The results presented in this study show that the modeling approach used is a fruitful way to model the coal-burning circulating fluidized bed process and to simulate its emissions.

The CFB boiler used as the case process is a coal-fired facility. In the future, it would be appropriate to test the methodology in other CFB processes incinerating different types of fuels, such as bark, biomass and even waste. This is because these inhomogeneous fuels are extremely challenging when it comes to efficient and stable combustion, while their use is also bound to increase in the future. Applying the method more widely would offer the opportunity to validate the method and make it a general approach for data-driven diagnostics and modeling of CFB boilers.

7. REFERENCES

1. J.A. Harding, M. Shahbaz, Srinivas, A. Kusiak. ”*Data Mining in Manufacturing: A Review*”. Journal of Manufacturing Science and Engineering, 128(4):969-976, 2006

2. S. Haykin, S. “*Neural Networks: A Comprehensive Foundation*”, Prentice Hall, Upper Saddle River, NJ (1999)

3. M.R.M Meireles, P.E.M Almeida, M.G. Simões. “*A Comprehensive Review for Industrial Applicability of Artificial Neural Networks*”. IEEE Trans. Industrial Electronics, 50(3):585-601, 2003

4. I.M. Mujtaba, M.A. Hussain (eds.). ”*Application of Neural Network and Other Learning Technologies in Process Engineering*”, Imperial College Press, London, UK (2001)

5. M. Heikkinen, T. Hiltunen, M. Liukkonen, A. Kettunen, R. Kuivalainen, Y. Hiltunen. ”*A Modelling and Optimization System for Fluidized Bed Power Plants*”. Expert Systems with Applications, 36(7):10274-10279, 2009

6. M. Heikkinen, T. Heikkinen, Y. Hiltunen. ”*Process States and Their Submodels Using Self-Organizing Maps in an Activated Sludge Treatment Plant*”. In Proc. 48th Scandinavian Conference on Simulation and Modeling (SIMS 2008) [CD-ROM]. Oslo University College, 2008

7. M. Heikkinen, V. Nurminen, T. Hiltunen, Y. Hiltunen. ”*A Modeling and Optimization Tool for the Expandable Polystyrene Batch Process*”. Chemical Product and Process Modeling, 3(1), Article 3, 2008

8. M. Heikkinen, A. Kettunen, E. Niemitalo, R. Kuivalainen, Y. Hiltunen. ”*SOM-based method for process state monitoring and optimization in fluidized bed energy plant*”. In W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (eds.), Lecture Notes in Computer Science 3696, pp. 409-414. Springer-Verlag Berlin Heidelberg (2005)

9. M. Liukkonen, M. Heikkinen, E. Hälikkä, R. Kuivalainen, Y. Hiltunen. ”*Emission Analysis of a Fluidized Bed Boiler by Using Self-Organizing Maps*”. In M. Kolehmainen, P. Toivanen, B.

Beliczynski (eds.), Lecture Notes in Computer Science 5495, pp. 119-129. Springer-Verlag Berlin Heidelberg (2009).

10. M. Liukkonen, T. Hiltunen, E. Havia, H. Leinonen, Y. Hiltunen. "*Modeling of Soldering Quality by Using Artificial Neural Networks*". IEEE Trans. Electronics Packaging Manufacturing, 32(2):89-96, 2009

11. M. Liukkonen, E. Havia, H. Leinonen, Y. Hiltunen. "*Application of Self-Organizing Maps in Analysis of Wave Soldering Process*". Expert Systems with Applications, 36(3P1):4604-4609, 2009

12. T. Kohonen. "*Self-Organizing Maps*", 3rd ed., Springer-Verlag, Berlin Heidelberg New York (2001)

13. R.D. Reed, R.J. Marks II. "*Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*", MIT Press, Cambridge, Massachusetts (1999)

14. M. Kasslin, J. Kangas, O. Simula. "*Process State Monitoring Using Self-Organizing Maps*". In I. Aleksander, J. Taylor (eds.), Artificial Neural Networks 2, Vol. I, pp. 1532-1534. North-Holland, Amsterdam, Netherlands (1992)

15. E. Alhoniemi, J. Hollmén, O. Simula, J. Vesanto. "*Process Monitoring and Modeling Using the Self-Organizing Map*". Integrated Computer Aided Engineering, 6(1):3-14, 1999

16. M. Liukkonen, M. Heikkinen, T. Hiltunen, E. Hälikkä, R. Kuivalainen, Y. Hiltunen. "*Modeling of Process States by Using Artificial Neural Networks in a Fluidized Bed Energy Plant*". In I. Troch, F. Breitenecker (eds.), Proc. MATHMOD 09 VIENNA, Full Papers Volume [CD-ROM]. ARGESIM – Publishing House, Vienna (2009)

17. J. MacQueen. "*Some methods for classification and analysis of multivariate observations*". In Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Statistics, pp. 281-297. University of California Press, Berkeley and Los Angeles (1967)

18. D.L. Davies, D.W. Bouldin. "*A Cluster Separation Measure*". IEEE Trans. Pattern Recognition and Machine Intelligence, 1(2):224-227, 1979

19. J. Freeman, D. Skapura. "*Neural Networks Algorithms, Application, and Programming Techniques*", Addison-Wesley Publishing Company, Menlo Park, California (1991)

20. R. Hecht-Nielsen. "*Neurocomputing*", Addison-Wesley Publishing, San Diego, California (1990)

21. P.J. Werbos. "*The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*", John Wiley & Sons, New York (1994)

22. C.J. Willmott. "*On the validation of models*". Physical Geography, 2:184-194, 1981

23. P. Basu. "*Combustion of coal in circulating fluidized-bed boilers: a review*". Chemical Engineering Science, 54(22):5547-5557, 1999

24. B. Leckner, A. Lyngfelt. "*Optimization of Emissions from Fluidized Bed Combustion of Coal, Biofuel and Waste*". International Journal of Energy Research 26(13):1191-1202, 2002

25. K. Svoboda, M. Pohořelý. "*Influence of operating conditions and coal properties on NO_x and N_2O emissions in pressurized fluidized bed combustion of subbituminous coals*". Fuel 83(7-8):1095-1103, 2004

26. A. Tourunen, J. Saastamoinen, H. Nevalainen. ”*Experimental Trends of NO in Circulating Fluidized Bed Combustion*”. Fuel 88(7):1333-1341, 2009

CALL FOR PAPERS

Journal: International Journal of Data Engineering (IJDE)
Volume: 1 **Issue:** 3
ISSN:2180-1274
URL: http://www.cscjournals.org/csc/description.php?JCode=IJDE

About IJDE

Data Engineering refers to the use of data engineering techniques and methodologies in the design, development and assessment of computer systems for different computing platforms and application environments. With the proliferation of the different forms of data and its rich semantics, the need for sophisticated techniques has resulted an in-depth content processing, engineering analysis, indexing, learning, mining, searching, management, and retrieval of data.

International Journal of Data Engineering (IJDE) is a peer reviewed scientific journal for sharing and exchanging research and results to problems encountered in today's data engineering societies. IJDE especially encourage submissions that make efforts (1) to expose practitioners to the most recent research results, tools, and practices in data engineering topics; (2) to raise awareness in the research community of the data engineering problems that arise in practice; (3) to promote the exchange of data & information engineering technologies and experiences among researchers and practitioners; and (4) to identify new issues and directions for future research and development in the data & information engineering fields. IJDE is a peer review journal that targets researchers and practitioners working on data engineering and data management.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE.

IJDE List of Topics

The realm of International Journal of Data Engineering (IJDE) extends, but not limited, to the following:

- Approximation and Uncertainty in Databases and Pro
- Autonomic Databases
- Data Engineering
- Data Engineering Algorithms
- Data Engineering for Ubiquitous Mobile Distributed
- Data Engineering Models
- Data Integration
- Data Mining and Knowledge Discovery

- Data Ontologies
- Data Query Optimization in Databases
- Data Warehousing
- Database User Interfaces and Information Visualiza
- Metadata Management and Semantic Interoperability
- Personalized Databases
- Scientific Biomedical and Other Advanced Database
- Social Information Management
- Data Privacy and Security
- Data Streams and Sensor Networks
- Database Tuning
- Knowledge Technologies
- OLAP and Data Grids
- Query Processing in Databases
- Semantic Web
- Spatial Temporal

Important Dates

Volume: 1
Issue: 3
Paper Submission: July 31, 2010
Author Notification: September 01, 2010
Issue Publication: September/October 2010

CALL FOR EDITORS/REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***International Journal of Data Engineering (IJDE)***. CSC Journals would like to invite interested candidates to join **IJDE** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org