Project Rough Draft Peer-Review Form(G02)

CSCI 8701 Overview of Database Research

Title of Paper: Imputation of Missing Values for Hierarchical Population data via Data Mining and Machine Learning

Authors: Muhammad Aurangzeb Ahmad, Nupur Bhatnagar

Reviewer Team (Name, Student Ids):
Prasad Sriram (sprasad@cs.umn.edu)
Nilu Thakur (nthakur@cs.umn.edu)

Date Review Completed: November 13, 2005.

SUMMARY:
FOCUS:
Does the paper clearly identify the problem it is addressing ?
-Yes, the paper clearly identifies the problem by giving a good example at the beginning.
-But the problem is not motivated. More examples should be given in order to stress the significance of this problem.

Does the paper clearly explain related work and their limitations?
-No, this part needs to be explained more in detail.
-There has been quite a lot of work on filling missing values in the database. The paper should describe more references and should clearly point out the reason of choosing Bayesian Belief Networks and Constrained Naïve Bayes.

Does the paper identify its key contributions?
Yes, the paper identifies its key contribution as using some unused machine learning techniques to find the missing values.

Does the paper present any evidence to support the contribution claim?
No, the paper doesn't present any validation methodology to substantiate their claims about increased accuracy. But the authors do mention their intention of using weka and other packages to conduct experiments.

TECHNICAL EVALUATION:
Is the literature survey complete?
-There is no list of references in the paper. They need to be cited in all places where related work is described inside the paper.
-Although, the authors describe Naïve Bayes technique in detail, they don't describe other techniques that have been used for this problem.

Is the work novel relative to the literature? Explain.
Yes, the authors seem to be attempting new machine learning techniques that have not been used before for this missing values problem.
As a reviewer do you agree with the contribution claims? Explain.
As the authors have not validated their approach with experiments, it is quite difficult to agree before looking at their actual results.

READABILITY AND ORGANIZATION:
Is the paper easy to read and understand to students in this course (CSCI 8701)?
-No, some parts of the paper are quite challenging to read as many parts of the proposed approach like relevant key equations about determining posterior probabilities have been left incomplete. Last line of data sets section is left incomplete.
-Constrained Naive Bayes approach is left incomplete. Should provide all the necessary equations related to determining mean and variance.
-The authors haven't described what is meant by hierarchical data. It would be better if the authors explain the term hierarchical in "hierarchical" census data.
-The related work section needs to be re-organized to make it more coherent.
-In the motivation section, it would be good if the authors explicitly mention the existing techniques rather than mentioning "various machine learning techniques". This would give the readers a fairly good idea of what to expect later in the paper.
-Should look into the formatting details. Margin space is too less.

Is the paper self-contained?
Yes, the paper's material can be understood by reading just the material present inside the paper.

Is the paper length reasonable?
Some sections of the paper for example the first two pages can be made more concise and clear. Otherwise, the length seems to be pretty reasonable.

Does it include sufficient number of figures and tables?
Yes, it includes tables wherever required.

STRENGTHS:
What are the strengths of this paper?
The paper attempts to apply different machine learning techniques with an objective to achieve higher accuracy than some other learning techniques.

AREAS FOR IMPROVEMENT:
How can this paper be improved? If you were to rewrite this paper, what revisions would you consider?
-I would recommend the authors to look into techniques like spectral analysis. These kinds of machine learning techniques are often employed to fill missing values in a matrix.
-The authors could add some sentences to say how privacy is preserved in their databases as it is a major concern.

-Ages cannot be the same in 1850-1870 census and in 1880. Should make appropriate changes in the numbers used. Also what if some person's records are not available in previous census records? A new child may have born. So do these proposed techniques give accurate results in these situations?
-In the constraints part of the problem definition, the constraints need to be explained in detail. Why is the ordering scheme of the household significantly important?
-Naïve Bayes is in deep trouble when the generative model assumptions of the data are violated. Also, Naïve bayes assumes total independence between variables. But the authors mention that some variables in their datasets are highly dependent. How is Naïve bayes expected to perform well as the data set doesn't meet those assumptions.
-While there are many graphical models available, why Bayesian Belief Networks and why not other graphical models that capture dependency?
-Also state what literature survey was performed in the context of these learning techniques?
-The fourth key assumption which says variables are assumed to be independent is not quite clear. Reading the paper gives an idea that variables are dependent in the data sets available.