Modified Deviation Approach to Deal with Missing Attribute values in Data Mining with Different percentage of Missing Values

Pallab Kumar Dey
Department Of Computer Science
The University of Burdwan
Bardhaman-713104,India.

Sripati Mukhopadhyay
Department Of Computer Science
The University of Burdwan
Bardhaman-713104,India.

ABSTRACT

Information System having missing attribute values (in practical) hampers accurate estimation of Data Mining. If missing attribute values can be predicted in the pre-processing stage of data mining then it will help to improve the accuracy, and the existing data mining algorithms can also be applied based on complete data. In this work different type of methods available to handle incomplete information system have been discussed, and there after an algorithm has been proposed by which missing attribute values may be replaced with minimum complexity. It is shown that proposed algorithm is better by applying it on different sets of data with different percentage of missing values.

General Terms:

Data Mining, Pre-processing

Keywords:

Data Mining ,Incomplete Information, Missing attribute Values , pre-processing , Modified Deviation approach

1. INTRODUCTION

Data Mining is a systematic approach for finding rules and pattern from large dataset. Real-life datasets may be incomplete for several reasons and Data mining with incomplete data is a challenging job. Various methods are available to deal with missing attribute values. To change a Incomplete Information System (i.e., datasets with missing attribute values) in a complete Information System objects, which have at least one missing value for an attribute, have to delete or ignore .But here it has been compromised with accuracy. So it can be applied only when dataset is too large and number of missing object is small. By list-wise or pair-wise objects may be deleted [1]. Knowledge can be extracted or rule can be generated directly from Incomplete Information [12, 7]. Here incompleteness of dataset have to manage at the time of rule generation.But these methods complexity have to consider,also here it's not possible to use available data mining algorithm. In pre-processing step if Incomplete Information can be translated into Complete information then available Data Mining algorithms can be used.
In preprocessing step missing value may be handled by different methodology. Missing values may be filled up the by maximum occurring attribute value or by maximum occurring attribute value within same concept(decision value)[2, 9].But here prediction error is not considerable.In preprocessing step missing value may be filled by all feasible domain values of the attribute or by all feasible domain values of same concept(decision value) of the attribute[6, 5].But it will increase number of entity which effect its complexity and accuracy.
In preprocessing step missing values may be treated by various statistical based methods [8, 4, 11, 13, 3] which are efficient and easy to implement in any software packages.In mean-mode method [8] missing values are replaced by mean of observe attribute value for numerical data and by maximum occurring attribute value for linguistic variable.In closest fit approach [4] missing values are replaced by mean of, observe attribute values and mean-of previous and successor values.Missing values may be replaced randomly such that standard deviations remain same[11] but complex to implement. In deviation approach[3] mean absolute deviation of observe values have been considered to fill missing attribute values.Mean-mode approach and closest fit approach are easy to implement but extreme point (from mean) will suffer. In Deviation approach extreme point can be better predicted but middle points where previous value and successor value are in same direction from mean will suffer.
In this paper a method has been proposed to deal with missing values in the category of MCAR(missing completely at random)[10] where all missing values can be predicted by same weightage. Here individual missing values(not only statistical computation result) may be estimated better than other methods. The proposed algorithm is also simple and low cost.

2. MATHEMATICAL AND COMPUTER MODELING

First observe mean ($\bar{A}_j$) for jth attribute may be calculated as follows-
$\bar{A}_j = \frac{1}{m}\sum_{i=1}^{m} V_{ij}$ where m is the number of non missing attribute value for jth attribute.

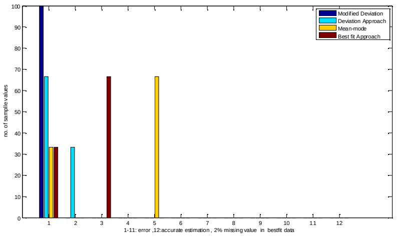
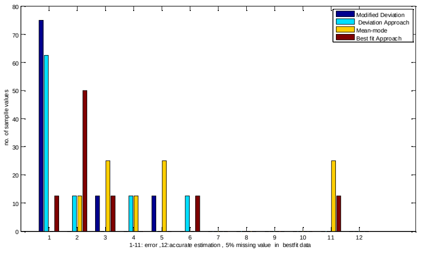
Now mean deviation ($\bar{A}_{jMD}$) based on not null Previous(V_{ijpre}) value of missing attribute value(V_{ij}) and not null successor(V_{ijflw}) value of missing attribute value from observe mean($\bar{A}_j$) may be calculated as follows-

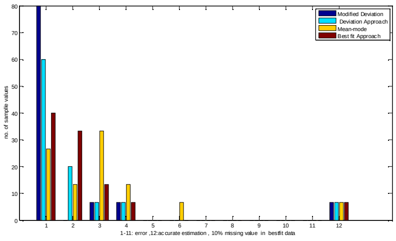
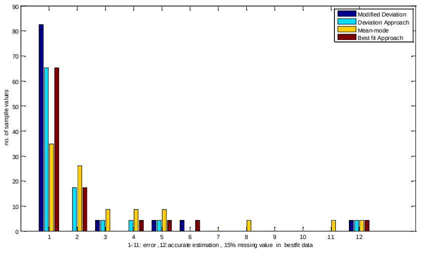
$$\bar{A}_{jMD} = \frac{1}{2}[|\bar{A}_j - V_{ijpre}| + |\bar{A}_j - V_{ijflw}|]$$

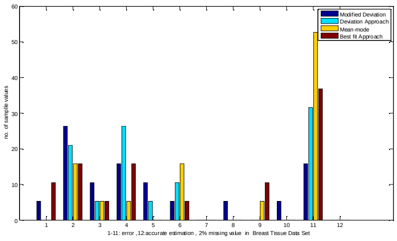
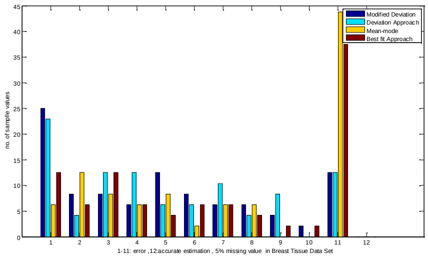
Mean of Previous (V_{ijpre}) value and successor (V_{ijflw}) value may be calculated as follows -

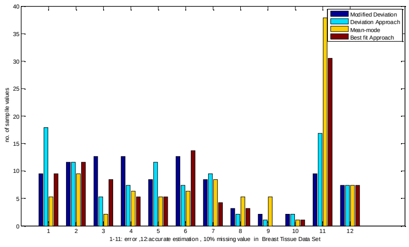
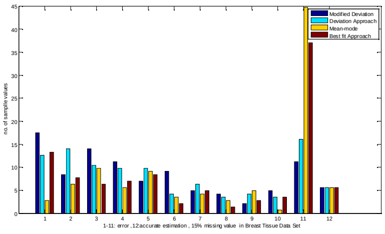
$$\bar{V}_{ijPF} = \frac{V_{ijpre}+V_{ijflw}}{2}$$

<table>
<tr><td colspan="3">Comparison of Different Methods with Different Data sets and Different percentage of Missing Attribute values
Row Denotes(Error percentage) ::
1(<=10%),2(>10%,<=20%),3(>20%,<=30%),4(>30%,<=40%),5(>40%,<=50%),6(>50%,<=60%),
7(>60%,<=70%),8(>70%,<=80%),9(>80%,<=90%),10(>9%,<=100%),11(>100%),12(AccuratePrediction)
Column Denotes:: no of sample values(percentage of total missing values)
So (X,Y)=(25,1) denotes, predication error is less than 10% in 25% missing sample values.</td></tr>
<tr><td>Best fit Data

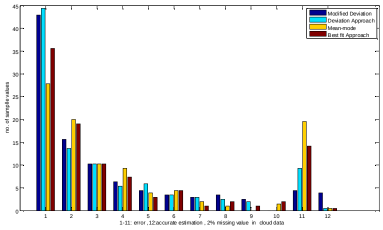
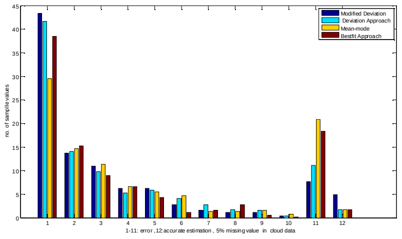
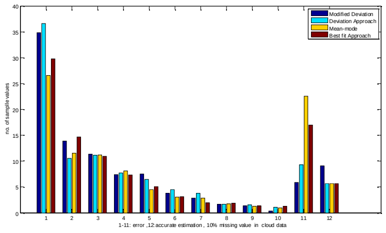
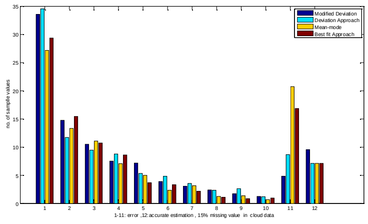
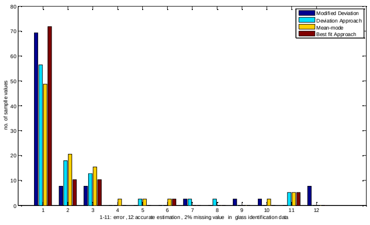
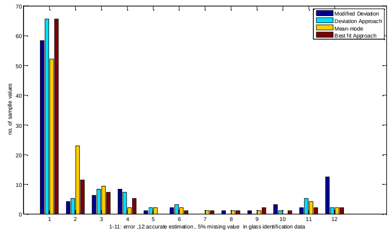
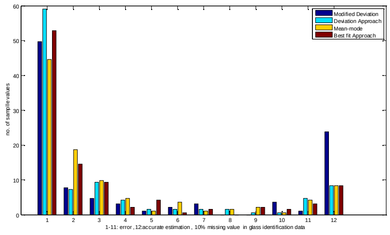
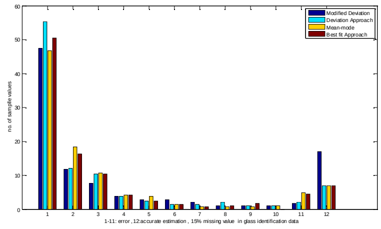
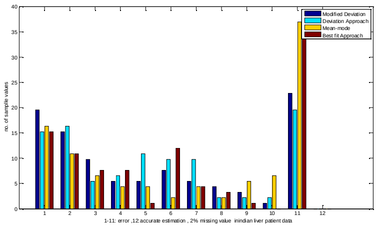
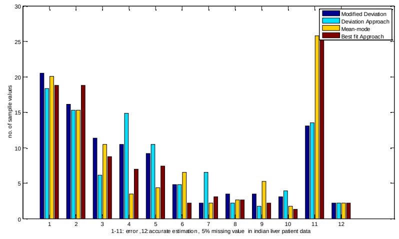
2% & 5% Misng Att val</td><td>
</td><td>
</td></tr>
<tr><td>Bestfit Data

10% & 15% Misng Att val</td><td>
</td><td>
</td></tr>
<tr><td>Breast Tissue Data Set

2% & 5% Misng Att val</td><td>
</td><td>
</td></tr>
<tr><td>Breast Tissue Data Set

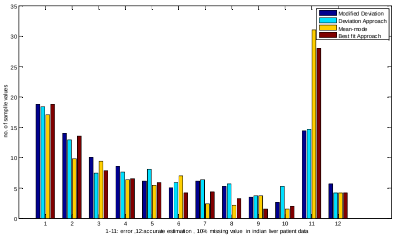
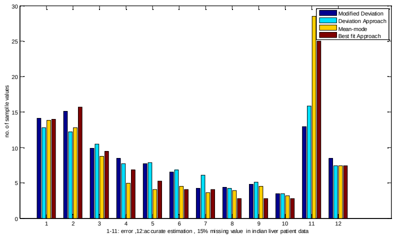
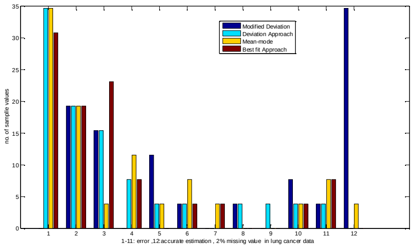
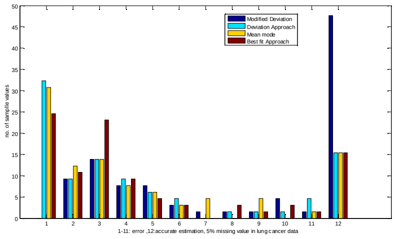
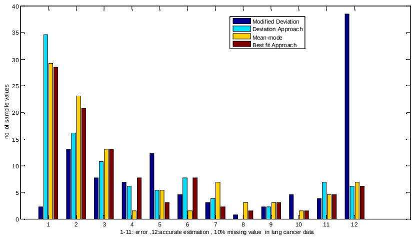
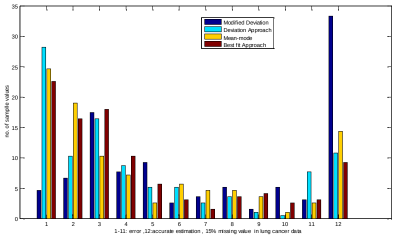
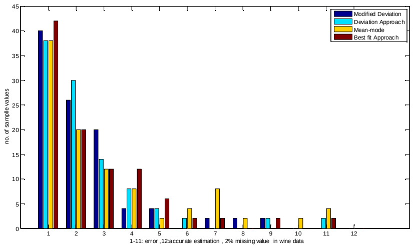
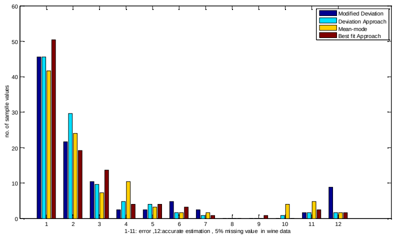
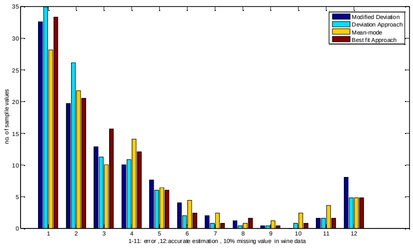
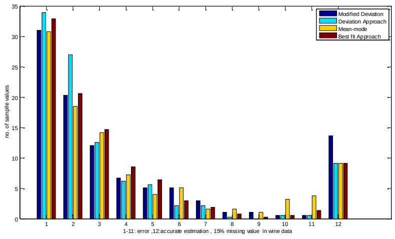
10% & 15% Misng Att val</td><td>
</td><td>
</td></tr>
</table>

Previous (V_{ijpre}) value and successor (V_{ijflw}) value have been taken as the estimator of missing value to calculate deviation direction. It may be assumed that missing value has positive deviation if previous value and successor value both have positive deviation from observe mean. so in that case :

Cloud data 2% & 5% Misng Att val		
Cloud data 10% & 15% Misng Att val		
glass identification 2% & 5% Misng Att val		
glass identification 10% & 15% Misng Att val		
indian liver patient 2% & 5% Misng Att val		

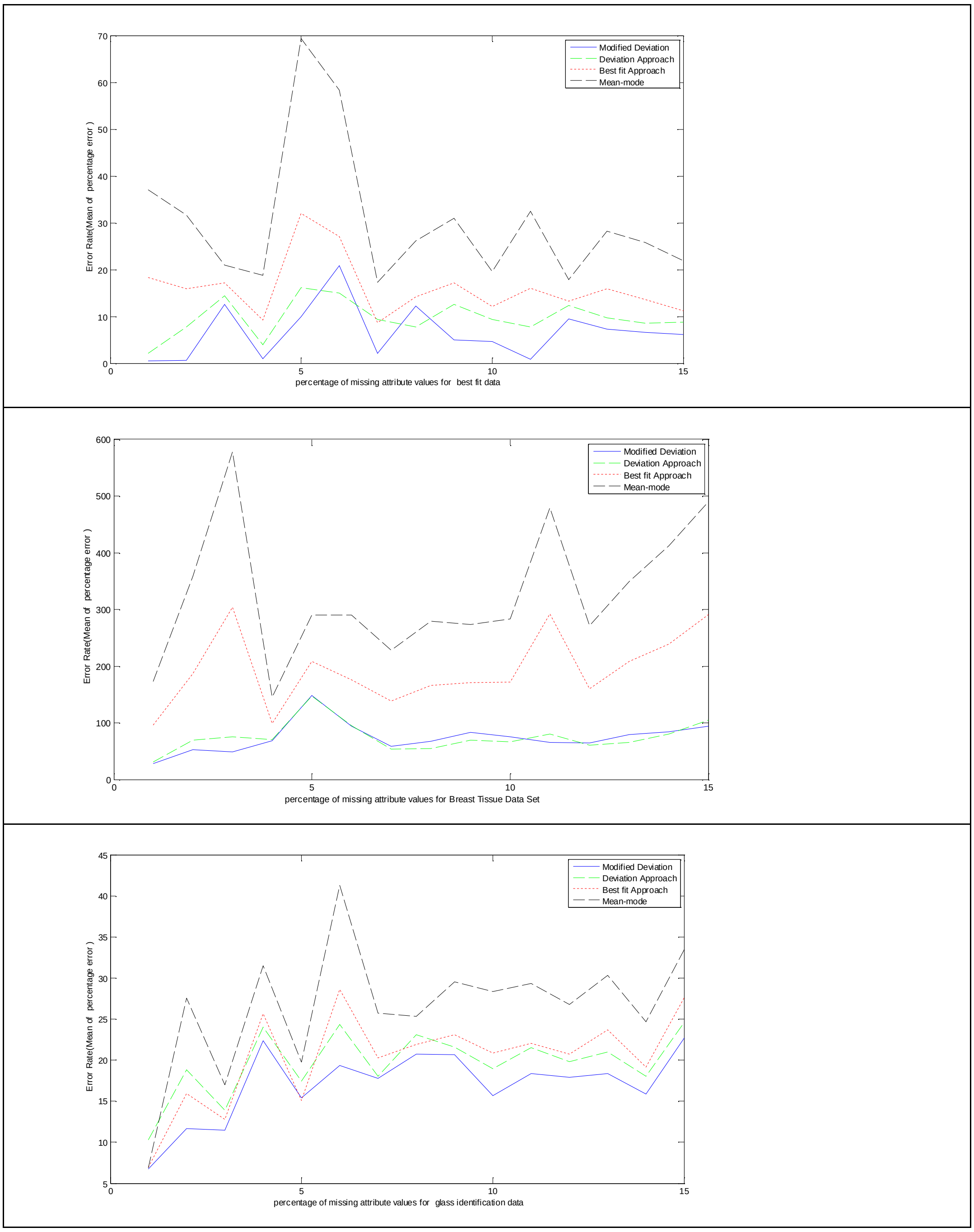
$$V_{ij} = \frac{\bar{V}_{ijPF}+(\bar{A}_j+\bar{A}_{jMD})}{2}$$

It may be assumed that missing value has negative deviation if previous value and successor value both have negative deviation from observe mean. so in that case :

indian liver patien t 10% & 15% Misng Att val		
lung cancer 2% & 5% Misng Att val		
lung cancer 10% & 15% Misng Att val		
Wine data 2% & 5% Misng Att val		
Wine data 10% & 15% Misng Att val		

$$V_{ij} = \frac{\bar{V}_{ijPF}+(\bar{A}_j-\bar{A}_{jMD})}{2}$$

It may be assumed that missing value has no deviation if previous value and successor value have deviation in opposite direction from observe mean. so in that case :



$$V_{ij} = \frac{\bar{V}_{ijPF}+\bar{A}_j}{2}$$

2.1 Algorithm

According to the above mathematical model the following algorithm has been proposed :

```
input  : Incomplete information System S, S= $A_j, V_{ij}$ :
         j=1,2,...,k; i=1,2,...,n where $V_{ij}$ may be missing
         k=number of Attributes, n=number of Objects
output: Complete Information System

for Each Attribute $j$ do
    $\bar{A}_j = \frac{1}{m}\sum_{i=1}^{m} V_{ij}$ ;
    for Each Attribute $i$ do
        if $V_{ij}$ missing then
            Find not null, previous value ($V_{ijpre}$ ) and
            successor value($V_{ijflw}$);
            $\bar{V}_{ijPF} = \frac{V_{ijpre}+V_{ijflw}}{2}$;
            $\bar{A}_{jMD} = \frac{1}{2}[|\bar{A}_j - V_{ijpre}| + |\bar{A}_j - V_{ijflw}|]$;
            if $V_{ijpre} \lneqq \bar{A}_j$ and $V_{ijflw} \lneqq \bar{A}_j$ then
                $V_{ij} = \frac{\bar{V}_{ijPF}+(\bar{A}_j-\bar{A}_{jMD})}{2}$;
            else if $V_{ijpre} \gneqq \bar{A}_j$ and $V_{ijflw} \gneqq \bar{A}_j$ then
                $V_{ij} = \frac{\bar{V}_{ijPF}+(\bar{A}_j+\bar{A}_{jMD})}{2}$
            end
            else
                $V_{ij} = \frac{\bar{V}_{ijPF}+\bar{A}_j}{2}$
            end
        end
    end
end
```

2.2 Analysis of Algorithm Complexity

Let k is the no. of attribute and n is the no. of Object present.Then first for loop will execute k times and second for loop will execute n times.So total time complexity for the above algorithm is $O(K) * O(n) = O(k * n)$.Clearly Space complexity is also O(1). So computational complexity for proposed algorithm is simple.

3. EXPERIMENTAL RESULT

Seven real data sets have been used where missing attribute values frequently occur.Six data sets accessible from http://archive.ics.uci.edu/ml/ and one data set taken from[4].Decision attributes and entity sets containing at least one missing attribute value have been deleted.In every data sets existing attribute values are randomly replaced by different percentage of symbolic lost values(missing values).Due to space limitation only few result of our experiments are presented in form of figures.These figure are self explanatory.From the figure it is clear that proposed algorithm can approximate missing value better than others.

4. CONCLUSIONS

Considering complexity the efficiency the proposed algorithm may be the best in certain field of application.In a large range of percentage of missing value this algorithm estimate consistent result. In this work application of proposed algorithm on numerical attribute values where missing data is MCAR have been discussed. Depending upon nature of missing attribute values and percentage of missing attribute value one method have to select to change it in a complete information system.

5. REFERENCES

[1] A. Acock. Working with missing values. *Journal of Marriage and Family*, 67:1012–1028, 2005.

[2] P. Clark and T. Niblett. *The CN2 induction algorithm,Machine Learning 3*. 1989.

[3] Pallab K. Dey and Sripati Mukhopadhyay. Deviation approach to missing attribute values in data mining. *International Journal of Advance Research in Computer Science*, 3(3):–, 2012.

[4] Sanjay Gaur and M.S. Dulawat. A closest fit approach to missing attribute values in data mining. *International Journal of Advances in Science and Technology*, 2(4):–, 2011.

[5] J. W. Grzymala-Busse. On the unknown attribute values in learning from examples. 542:368–377, 1991.

[6] J. W. Grzymala-Busse and Hu Ming. A comparison of several approaches to missing attribute values in data mining. pages 378–385, 2001.

[7] J. W. Grzymala-Busse and A. Y. Wang. Modified algorithms lem1 and lem2 for rule induction from data with missing attribute values. In Research Triangle Park, editor, *Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97), Third Joint Conference on Information Sciences (JCIS'97)*, page 6972, 2–5 1997.

[8] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, 2001.

[9] Bratko I. Knonenko and I. Roskar. *E.: Experiments in automatic learning of medical diagnostic rules.* Technical Report, Jozef Stefan Institute, Lljubljana, Yugoslavia, 1984.

[10] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, 1987.

[11] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, 1999. hardcopy.

[12] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[13] G.Weckman W.Young and W.Holland. A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. *Theoretical Issues in Ergonomics Science*, 12(1):15–43, 2011.