“Make decisions even with incomplete info. You’ll never have all the info you need. What matters is what you do with the info you have.”

— Ziad K. Abdelnour, Economic Warfare: Secrets of Wealth Creation in the Age of Welfare Politics
CHAPTER 1
INTRODUCTION

Making decision processes are highly dependent on the availability of data, from which information can be extracted. All engineering, business and economic decisions are somehow related to the information available at the time of creating such decisions. It is for this reason that the problem of missing data affects a variety of research and application areas in fields such as scientific, engineering, economics, finance and many more. Most predictive and decision making models designed to use a specified number of values will breakdown when one or more values are not available.

Information plays a very important role in our life. Advances in many research field depend on the ability of discovering knowledge in very large data bases. A lot of businesses base their success on the availability of marketing information. This kind of data is usually big, and not always easy to manage.

1.1 Scope of work and problem definition

Scientists from different research areas have developed methods to analyze huge amounts of data and to extract useful information. These methods may extract different kinds of knowledge, depending on the data and on user requirements. In particular, one important knowledge discovery task is supervised learning. Today, there exist many methods to build classifiers, belonging to different fields, such as artificial intelligence, soft computing, and statistics.

Unfortunately, traditional methods usually cannot deal directly with real-world data, because of missing or wrong items. This thesis concerns the former problems; the unavailability of some values. The majority of interesting data bases is incomplete, i.e., one or more values are missing inside some records, or some records are missing at all. There exist many techniques to manage data with missing items, but no one is absolutely better than the others. Different situations require different solutions. As Allison says, “the only really good solution to the missing data problem is not to have any”.

There are some methods already used for handling missing data such as, in expectation-maximization is an approach in which the value of statistics which would be computed if a complete dataset were available is estimated, taking in to the account the pattern of missing data. Value of individual missing data items are not imputed (estimated). In list-wise deletion method all cases with missing attribute values are deleted from the dataset. For example: Five individuals have subjects in group, missing one or more variable, and then omit those individuals from the analysis.

In Pair-wise deletion approach, each element of the inter correlation matrix is estimated using all available data. For Example, if one participant reports his income and life satisfaction index, but not his age, he is included in the correlation of income and life satisfaction, but not in the correlation involving age. Using these cases, these selected variables with complete data can be
analyzed or use all available data to estimate parameters of the model. It is also called available case analysis.

Full analysis takes full account of all information available, without the distortion resulting from using inputted values.

Due to the availability of software and the advancement of computational power, maximum likelihood and multiple imputations (Royston, P. 2009)\textsuperscript{124} and as well as Rough Set (Pawlak, Z. 1991)\textsuperscript{104} and Soft Set (Molodtsov, D. 1999)\textsuperscript{90} have been introduced as solutions to the incomplete data problem. Although these methods have given considerable results in this domain, the impact that the data and missing data imputation has no decision making has, until recently, not been assessed. This thesis contributes to this knowledge by first introducing a new computational intelligent model that integrates Rough Set and Soft Set Theory to impute missing data. This work has shown that although missing data and the imputation thereof has an effect on decision making, the degree of that effect is dependent on the decision making framework and on the model used for data imputation.

1.2 Objectives

In real datasets, like, e.g., surveys and clinical trials, it is quite common to have observations with missing values for one or more input features. The first issue in dealing with the problem is determining whether the missing data mechanism has distorted the observed data.

(Little and Rubin 2002)\textsuperscript{92} and (Rubin 2003)\textsuperscript{128} distinguish between basically three missing data mechanisms these are MAR, MCAR and MNAR. Data are said to be missing at random (MAR) if the mechanism resulting in its omission is independent of its (unobserved) value. If its omission is also independent of the observed values, than the missingness process is said to be missing completely at random (MCAR). In any other case the process is missing not at random (MNAR), i.e., the missingness process depends on the unobserved values.

1.3 Contributions

The main contributions of the thesis are:

- The issues of Missing Data and find the suitability of rough set approach, a soft computing technique.

- Try to automate the transformation of imprecise data into knowledge using the rich rough set theory.

- Try to eliminate any drawbacks available in these methodologies either giving a modified algorithms or propose a new algorithm based on the feasibility of the datasets and to handle the missing attribute values or ill known data or missing data under imprecise and/ or uncertainty for retrieving information.

- In these methods of handling missing attribute values in data mining are described. These methods are categorized into sequential and parallel. In sequential methods, missing attribute values are imputed by known values first, as a pre processing, then the knowledge
is acquired for a data set with all known attribute values. In parallel methods, there is no pre processing, i.e., knowledge is acquired directly from the original data sets.

- Emphasis is put on rule induction. Methods of handling attribute input values for decision tree generation are only briefly discussed.

1.4 Organization of the Thesis

Chapter 1 starts with an introduction of the work carried out in this thesis.

The outline of the rest of the thesis is as follows:

Chapter 2 comprises the literature survey in Missing Data. Research in a number of academic fields has shown that missing data operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their targets.

Chapter 3 introduces the concepts of Rough set, fuzzy set and soft set theories in more detail. In this correlate the discussed missing data techniques of previous chapter and try to find out their algorithmic suitability in comparing to the model functional analysis.

In Chapter 4, Rough Set Approach to Rule Induction is discussed and it handles the Incomplete Data, is studied.

Chapter 5 Focus on Rough Set Rule Based Technique for the Retrieval of Missing Data in Malaria Diseases Diagnosis, where a real data set is analysed and missing data are retrieved.

Chapter 6 focuses missing data with Covering Based Rough Set and Soft Set. An analysis is carried out a malaria data set on these three approaches.

In Chapter 7, the various techniques like, Rough sets and Soft set analysed for data like student data set.

Chapter 8 discusses the Expert systems and incomplete information with architecture. We propose the rules and methods to handle uncertainty and vague information to solve the problem of missing data.

Finally, in Chapter 9 we summarize the scope and contributions of this project and discuss various avenues of future research work, which expand and complement the research presented in this project.