To

My beloved father

Janab Zubair Ahmad Saheb
ABSTRACT

This thesis addresses an in-depth investigation into the issues related to the construction of Knowledge Discovery in Databases (KDD) System for discovery of Quantified Hierarchical Censored Production Rules by using data mining techniques. The motivation for the work was the need for an alternative to the slow and expensive manual knowledge acquisition process in which a knowledge engineer interviews an expert repeatedly to acquire knowledge in the form of Quantified Hierarchical Censored Production Rules. Therefore the main aims were to develop a KDD system for anyone who is interested to:

- Quantify uncertainty in the data,
- Capture exceptions in the data and
- Needs knowledge in the form of Hierarchical structure.

The purpose of this study was to investigate whether data mining techniques could be used to discover such type of knowledge automatically from the large data sets.

In this context, special attention was paid to the methods of uncertainty quantification especially to Dempster Shafer Theory. Many Rule discovery algorithms and Machine learning methods were researched.

Production rules are one of the most popular and widely used knowledge representation languages. Early expert systems used production rules as their main knowledge representation language. For example, MYCIN, which is also considered one of the first research works in medical informatics, has production rules as its knowledge representation language. Most of the KDD approaches have used standard production rules of the form IF <condition> THEN <action> as the underlying knowledge representation. The standard production
rules have rigidity in handling uncertainty and exceptions. Production rules do not exhibit variable precision.

Production rule mining is very strong technique for data mining but with drawback of not including exceptions/ uncertainties existing in real life situations and rigidity in its syntax for not being able to incorporate hierarchies. Exceptions, which focus on a very small portion of a data set, have been ignored or discarded as noise in machine learning, but the goal of KDD is broader and it is always interesting to discover exceptions, as they challenge the existing knowledge and often lead to the growth of knowledge in new directions.

As an extension of production rules, Michalski and Winston proposed Censored Production Rule (CPR) of the form:

\[
\text{IF } \text{<condition> } \text{THEN } \text{<action> } \text{UNLESS } \text{C } \text{<censor>}
\]

Where C is the exception condition.

To address various problems and shortcomings of CPRs system, Bharadwaj and Jain have introduced a concept of Hierarchical Censored Production Rule (HCPR), by merging the concept of Hierarchical Production Rules with Censored Production Rules. HCPRs system is an underlying representational and computational scheme to enable logic based systems to exhibit variable precision in which certainty varies while specificity stays constant. Hence it is not able to impart control over the specificity part of a precision in decision-making. General form of HCPR is:

\[
\text{<Decision (if precondition)} \quad \text{(UNLESS censor conditions)} \quad \text{(GENERALITY general_information)}
\]

HCPRs are used to handle trade-off between the precision of an inference and its computational efficiency leading to trade-off between the certainty of a conclusion and its specificity.
As a special case (dropping the UNLESS operator) Hierarchical Production Rule (HPR) takes the form:

**Decision if <precondition>**

**GENERALITY <general_information>**

**SPECIFICITY <specific_information>**

The best way to express knowledge is in the form of hierarchies. Concept of Hierarchies organizes data and concepts in hierarchical forms, which helps expressing knowledge and data relationships in concise, high level terms and thus plays an important role in knowledge discovery process. Automatic generation of hierarchies can be a post-processing step, but integrating hierarchy generation process into the data-mining algorithm can dramatically reduce the execution time.

Automated discovery of Quantified hierarchical structure from large database plays fundamentally important role in data mining because it provides comprehensible results that capture real life inheritance of objects. Several efforts have been made in the recent past towards automated discovery of hierarchical structure in large databases.

The broad aim of the submitted research work is to construct a KDD system from large data sets based on inductive learning for automated discovery of quantified (DST/FUZZY LOGIC/ CF):

- Censor Production Rules
- Hierarchical Production Rules
- Hierarchical Censored Production Rules

The organization of this thesis is as follows:

Chapter 1 deals with introduction of our research area, problem statement and our research contribution.
Chapter 2 deals with different types of uncertainty and methods of its quantification, especially Dempster Shafer theory in detail.

Chapter 3 deals with related work in exception rule mining, uncertainty quantification and discovery of knowledge in hierarchical form.

Chapter 4 deals with overview of different types of knowledge representations used in the proposed work.

Chapter 5 deals with proposed framework of Data mining system for uncertainty quantification, from our research paper V.

Chapter 6 deals with two proposed approaches: Discovery of Quantified Censored Production Rules and Discovery of fuzzy CPR, from our research paper I and III.

Chapter 7 deals with proposed approach for discovery of Quantified Hierarchical Production Rules, from our research paper IV.

Chapter 8 deals with proposed approach for the Discovery of Quantified Hierarchical Censored Production Rules, from our research paper VI.

Chapter 9 deals with Conclusion, Summary and future work.

References deal with list of references used throughout our research work.

List of publication deals with information about our original research paper publication.