9. CONCLUSION

Data mining and reasoning under uncertainty are two very different disciplines of research and have very different roots. Uncertain reasoning is a sub area of Artificial Intelligence. These two fields share several common interests. Data mining methods are designed to find regularity in databases. They achieve this regularity by either removing uncertainty from data, or maintaining what is certain in the data. In addition, an interesting challenge is to utilize data mining techniques with missing and incomplete data. In many applications, there is a need to find unknown and interesting patterns even though the data may contain major flaws such as missing or inconsistent data values. The study of uncertainty can be viewed as a study of discovering hidden “true” system.

The proposed research work is the construction of KDD system, which is capable of providing Quantified Hierarchical Censored Production Rules. In this system uncertainty has been quantified through Dempster Shafer theory. The benefit of the work is that uncertain reasoning can be applied on the discovered final output in the following manner.

**Uncertain Reasoning in HCPRs System**

There are at least two basic problems when reasoning with uncertain facts or rules; namely the combination problem and the propagation problem. The combination problem refers to the aggregation of uncertain pieces of information issued from different sources dealing with the same matter. The propagation problem deal with the aggregation of the uncertainty concerning the satisfaction of the condition part of a rule with the uncertainty of the rule itself in order to deduce the uncertainty of the rule itself in order to deduce the uncertainty pervading the conclusion of the rule.
Beliefs are propagated across the rules. In this section the combination and propagation formulae for HCPRs system are discussed.

When censors are not under consideration

For the rule $P \rightarrow \neg D: \gamma$, $\delta$, the resulting expressions for support and plausibility of the decision are

$$s(D) = s(P) \ast s(P \rightarrow \neg D)$$
$$p(D) = 1 - s(P) \ast s(P \rightarrow \neg \neg D)$$

this implies

$$s(D) = s(P) \ast \gamma \quad (9.1)$$
$$p(D) = 1 - s(P) \ast \delta \quad (9.2)$$

When censors are under consideration

For the rule $D: \neg P \mid C$ the propagation formulae using $\alpha$, $\beta$, values are

$$s(D) = s(P) \ast (1 - p(C) \ast \alpha) \quad (9.3)$$
$$p(D) = 1 - s(P) \ast s(C) \ast \beta \quad (9.4)$$

The support and plausibility values represent the constraints that the censor value puts on the range of values that the decision may assume. The interval represented by these values is called possible range of the decision and denoted by $[s_{\text{poss}}, p_{\text{poss}}]$. The propagation formulae for the possible range are

$$s_{\text{poss}}(D) = s_{\text{poss}}(P) \ast (1 - p_{\text{poss}}(C) \ast \alpha) \quad (9.5)$$
$$p_{\text{poss}}(D) = 1 - s_{\text{poss}}(P) \ast s_{\text{poss}}(C) \ast \beta \quad (9.6)$$

where the reference to the possible ranges on the RHS of (9.5) and (9.6) allow the values to the propagated from rule to rule.

The possible range can leave some uncertainty concerning the value of the decision. The amount of uncertainty is equal to the width of the probability interval, and this uncertainty represents the probability mass which could not be assigned exactly to $D$ or $\neg D$. In other words, it is free to move between the two values. The $\gamma$ and $\delta$ values can be used to apportion this remaining probability mass. The $\gamma$ and $\delta$ values represent the probability of $D$ given $P$
when the censor is unknown. When the censor value is completely unknown, they are used to distribute the entire probability mass. When the censor value is partially known they can be used to distribute that portion of the probability mass which the censor value has failed to constrain. The range produced by use of $\gamma$ and $\delta$ value is called the most likely range and is denoted by $[s_{mi}, p_{mi}]$. The corresponding propagation formulae are:

$$s_{mi}(D) = s_{poss}(D) + [s_{mi}(P) - s_{mi}(C)]$$

$$p_{mi}(D) = p_{poss}(D) - [s_{mi}(P) - s_{mi}(C)]$$

Note that the propagation formulae for rules without censors are just the special case of (9.7) and (9.8), where

$$s_{poss}(D) = 0, p_{poss}(D) = 1$$

The most likely ranges of $P$ and $C$ appear on the RHS of (9.7) and (9.8) to allow the most likely range to propagate from rule to rule.

Possible and most likely formulae for HCPRs indicating propagation from level (i-1) --- level (i)

When censor is known

$$[s_{poss}(D)]_i = \min \{[s_{poss}(D)]_{i-1}, [s_{poss}(P)]_i\} * \{1 - [p_{poss}(C)]_i\} * a_i$$

$$[p_{poss}(D)]_i = 1 - \min \{[s_{poss}(D)]_{i-1}, [s_{poss}(P)]_i\} * \{s_{poss}(C)]_i\} * \beta_i$$

$$[s_{poss}(D)]_i = [s_{poss}(D)]_{i-1} + \min \{[s_{mi}(D)]_{i-1}, [s_{mi}(P)]_i\} * \delta_i * \{[p_{mi}(C) - s_{mi}(C)]\}$$

$$[p_{mi}(D)]_i = [p_{poss}(D)]_{i-1} - \min \{[s_{mi}(D)]_{i-1}, [s_{mi}(P)]_i\} * \gamma_i * \{[p_{mi}(C) - s_{mi}(C)]\}$$

When censor is unknown

$$[s_{poss}(D)]_i = 0$$

$$[p_{poss}(D)]_i = 1$$

$$[s_{mi}(D)]_i = \min \{[s_{mi}(D)]_{i-1}, [s_{mi}(P)]_i\} * \gamma_i$$

$$[p_{mi}(D)]_i = 1 - \min \{[s_{mi}(D)]_{i-1}, [s_{mi}(P)]_i\} * \delta_i$$
Where subscripts $i$ and $(i-1)$ indicate the $i^{\text{th}}$ level and the $(i-1)^{\text{th}}$ level respectively in HCPR-tree.

The final conclusion of the work is that the Quantified Rules with exceptions and Hierarchy have been discovered through data mining techniques and the output can be utilized for uncertain reasoning, which is a sub area of Artificial Intelligence.

9.1 **Summary**

In this thesis a KDD system for automated discovery of quantified production rules with exceptions and Hierarchy is presented. This work is done in three phases:

- Automated Discovery of Quantified Censor Production Rules
- Automated Discovery of Quantified Hierarchical Production Rules
- Automated Discovery of Quantified Hierarchical Censored Production Rules

With this thesis mainly we have developed four approaches:

**Proposed approach for D-S-Uncertainty Miner:** A data mining system that captures uncertainty in the data.

5. Post processing approach for discovery of Quantified Censored Production Rules.


7. Approach for the discovery of Quantified Hierarchical Censored Production Rules, using proposed Fusion algorithm between two knowledge bases.
In above four cases Dempster Shafer theory has been used as a method of uncertainty quantification.

The present study consists the introduction of six original publications (from I to VI). The author was the main contributor and responsible one, which were prepared in collaboration with the co-authors. Paper (I) presents proposed approach for Discovery of Quantified Censored Production Rules from the large set of Discovered Rules. Paper (II) presents a survey paper titled “A KDD Tool for Automated Discovery of Knowledge”. Paper (III) describes an alternate approach for “Discovery of Fuzzy Censored Production Rules from Large Set of Discovered Fuzzy If Then Rules”. Paper (IV) presents an approach to discover Quantified Hierarchical Production Rules. Paper (V) describes an approach for a KDD system that works as Uncertainty miner, using Dempster Shafer Theory. The last Paper (VI) presents new algorithm “Fusion” which discovers required final output Quantified Hierarchical Censored Production Rules from large data sets.

The defence of this thesis is organized as follows: Chapter 2 and chapter 3 provide background and motivation for this research. These two chapters provide definitions of important terms and describe related work. Chapter 4 provides different types of knowledge representations QCPR, QHPR and QHCPR: these are required output of the proposed KDD system. Chapter 5 describes a new approach of KDD system, which works as general uncertainty miner, by using Dempster Shafer theory. Chapter 6 presents first part of the research work, proposed approach for that discovery of Quantified Censored Production Rules from large data set. This is followed by another approach for the discovery of fuzzy Censored Production Rules. Chapter 7 presents second part of our research work, an approach for discovery of Quantified Hierarchical Production Rules. Chapter 8 describes last part of our
work, proposed Fusion algorithm for discovery of Quantified Hierarchical Censored Production Rules. This algorithm is applied between two knowledge bases QCPR and HPR and it produces required result of our work, QHCPR knowledgebase. Finally Chapters 9 concludes and details opportunities for further research.

9.2 Future Work

The proposed approach was applied on training data sets containing categorical and non-categorical attributes; so further work might consider training data set containing discrete value attributes. Another direction is the discovery of Fuzzy Hierarchical Censored Production Rules from large data set.