Abstract

The availability of very large volumes of such data has created a problem of how to extract from them useful, task-oriented knowledge. An enormous proliferation of databases in almost every area of human endeavor has created a great demand for new, powerful tools for turning data into useful, task-oriented knowledge. In efforts to satisfy this need, researchers have been exploring ideas and methods developed in machine learning, pattern recognition, statistical data analysis, data visualization, neural nets, etc. These efforts have led to the emergence of a new research area, frequently called data mining and knowledge discovery.

The most predominant representation of the discovered knowledge is the standard Production Rules (PR). However, standard production rules cannot handle incomplete and imprecise knowledge. To capture the imprecise knowledge about the real world Michalski & Winston have suggested the Censored Production Rule (CPR) as underlying representational that exhibit variable precision and supports an efficient mechanism for handling exceptions. A CPR is an augmented production rule of the form:

\[
\text{If } \langle \text{Condition}\rangle \text{ Then } \langle \text{Decision}\rangle \text{ Unless } \langle \text{Censor}\rangle
\]

where C (Censor) is an exception to the rule.

A CPR having more than one censor conditions, say, C₁, C₂, ..., Cₙ is called a Multiple Censored Production Rule (MCPR) denoted as

\[
\text{If } \langle \text{Condition}\rangle \text{ Then } \langle \text{Decision}\rangle \text{ Unless } (C₁ \lor C₂ \lor ... \lor Cₙ)
\]

Another alternative for handling exceptions is Ripple-Down Rules (RDRs)- a machine learning technique proposed by Compton and Jansen that represents the exception rules in a concise manner in order to keep all general rules consistent. The structural form of RDR is:

\[
\text{If } \langle \text{condition}\rangle \text{ Then } \langle \text{conclusion}\rangle \text{ Except If...}
\]

\[
\text{Else If...}
\]

A CPR is unable to capture the hierarchical structure inherent in the knowledge about the real world and hence would not impart any control over the specificity part of
precision in decision making. As an extension of CPR, the Hierarchical Censored Production Rules (HCPRs) system of knowledge representation, proposed by Bharadwaj & Jain (1992), exhibits both variable certainty as well as variable specificity and offers mechanisms for handling the trade-off between the two. A HCPR is of the form

\[
\begin{align*}
\text{Decision} & \text{ If } \langle \text{condition} \rangle \\
& \text{ Unless } \langle \text{censor} \rangle \\
& \text{ Generality } \langle \text{general info} \rangle \\
& \text{ Specificity } \langle \text{specific info} \rangle 
\end{align*}
\]

The main objective of this thesis is to discover Hierarchical Production Rules with Exceptions from large datasets. The underlying representation of the discovered knowledge is based on the Hierarchical Censored Production Rules (HCPRs) and proposed extensions of HCPRs and RDRs.

An automated discovery of Multiple Censored Production Rules (ADMCPRs) algorithm is developed as an extension of the Inductive Learning Algorithm (ILA) to discover any or all standard Production Rules (PRs)/ CPRs / MCPRs from set of examples.

In order to summarize large datasets and minimize the number of dataset accesses, a concept of Frequency matrix (Freq) is introduced.

Discovery of Production Rules with Fuzzy Hierarchy (DPRFH) algorithm is proposed using a novel approach, based on Freq matrix, that integrates the process of hierarchy generation and discovery of rules of the form:

\[ P \rightarrow D_k \]

\[
\begin{align*}
& \text{Generality [general classes]} \\
& \text{Specificity } [D_{k1}(d_1), \ldots, D_{ki}(d_i), \ldots, D_{kj}(d_j)]
\end{align*}
\]

where P is the set of preconditions and the specificity element D_{ki}(d_i) means that D_{ki} is a specific class of D_k with degree of subsumption d_i.
Further, the Discovery of Production Rules with Exceptions and Fuzzy Hierarchy (DPRFH) algorithm is presented as an extension of the DPRFH-algorithm. The extended algorithm discovers hierarchical rules with exceptions of the form:

\[ P \rightarrow D_k \]

Unless \[ [e_{k1} \lor \ldots \lor e_{ki} \lor \ldots] \]

Generality \{general classes\}

Specificity \{D_{k1}(d_1), \ldots, D_{kj}(d_i), \ldots, D_{kj}(d_j)\}\]

where \( e_{ki} \) is the \( i \)-th exception condition to the class \( D_k \).

In order to incorporate hierarchical structure into Ripple-Down Rules (RDRs), we have proposed a concept of Hierarchical Ripple-Down Rule (HRDR)- RDR augmented with Generality and Specificity operators. An HRDR-Discovery algorithm is proposed for the automated discovery of HRDR.

Experimental results are presented to demonstrate the effectiveness of the proposed algorithms.