CHAPTER 6

CONCLUSION AND SCOPE FOR FURTHER WORK

In this research work, a hybrid method for attribute selection is explored. The framework focuses on a combination of filter and wrapper methods to categorize and identify attributes. The main research findings are:

I. A combination of multiple methods for ranking (identifying the importance of an attribute) is always preferable than a single measure. The multiple method combination is more effective if they are derived independent of each other and belong to different categories (for e.g. Principal Component Analysis and Shannon’s entropy). The Principal Component Analysis stage and Shannon’s entropy account for the data dependent and independent method classes.

II. Attribute selection systems can benefit from a combination of ranking, clustering and correlation methods. The resultant combination needs adaptive learning systems for proper selection. This is very important as the method must not lose the key metrics – information gain, correlation, pair wise correlation and clustering tendency.
III. Methodologies that consider the correlation of attributes perform well. A good attribute subset is one that considers not just the best attributes, but attributes that are related to other attributes.

IV. Methodologies must account for differences in class distribution and sizes of the datasets. This is important as the algorithms must not be biased in any way and need solutions that can adapt to the sizes of the datasets.

**Scope for Further Work**

I. **Use of domain dependency in attribute selection:** This research work can be further extended by applying domain dependent knowledge. If any knowledge about the target domain is available, the method can benefit from ways to encode this information in the learning algorithm.

II. **Time optimization in attribute selection:** In this research work, the focus has been on getting the proper attribute subset rather than optimizing the time taken. Thus the time taken for attribute selection has not been optimized. Suitable algorithms can be developed or integrated with the existing approach for optimizing time.
III. **Attribute selection on object oriented dataset:** The dataset/data structure used in the research work can be further redesigned by considering the object oriented concepts. This leads to better time optimization in attribute selection.

IV. **Incorporation of other methods/models for attribute selection:** In this work, the aggregation process is stopped with limited iteration of attribute subset methods such as information gain, correlation and clustering. This can be further extended to other methods such as association, prediction analysis etc., to perform deeper analysis on attribute selection.

V. **Extension to other mining models:** This research work considers the relational dataset which contains alpha numerals for the attribute selection. More types of data such as multimedia, text, web usage, etc., can be considered by applying suitable mining models such as audio mining, video mining, text mining and web mining.
Contribution of Thesis

This research work has outlined an attribute selection algorithm that is adaptive and makes use of the best of clustering, correlation and ranking systems to form a hybrid system.

I. The work combines both filter and wrapper methods and has bench marked the results for a range of datasets to test the efficiency.

II. The work emphasizes a strategy to combine information gain, clustering and correlation aspects to form the best attribute subset.

III. A prototype for an adaptive hybrid attribute selection algorithm which utilizes all the attributes, static and dynamic, to make target identification more accurate and robust than using each type of attributes alone is developed.

IV. The developed method advocates an adaptive aggregation strategy using
   a. the gain ratio for candidate attributes,
   b. clustering methods to find the distribution of candidate attributes and
   c. correlation based strategies which measure the dependencies of different attributes.
V. The resultant adaptive method is implemented for the datasets from the UCI Machine Learning Repository and the results are correlated.

VI. The missing values on the selected subset of attributes are imputed with C4.5 and K-Means algorithm and the results are correlated for better classification accuracy.

The contribution of the thesis lies in multiple ways: choice of algorithms to get the measure of good attributes, examine their clustering and correlation tendencies and adapt the attribute aggregation methods depending on various factors (size and content of dataset).