Chapter 7

Conclusions & Future Work

In data mining, a diversity of concepts that support the knowledge discovery process has been encountered. Among them classification is being the most dominant category in the development of knowledge discovery model. To deal with huge databases, several fundamental ways are there to build the classifier and to extract the classification rules. Discretization is one of the most important, and often required preprocessing methods as it reduces the complexity of the data for the task of classification and improves the classification accuracy. The discretized data may be further processed by a pattern selection algorithm to improve the performance of a classification process through a consideration of only the most informative data. In a diversity of classifiers, recently neural networks are being able to model specific regions of the brain very accurately and to classify a variety of tasks outside the modeling realm.

The performance of the classifier is mainly linked with its ability of generalization. High generalization stipulates that the classifier can perform well on a new
data. The classifier with high generalization abilities can be built by using a neural network with optimal topology. The optimal topology of neural networks can be determined by the process of pruning. Pruning attempts to find a quick solution to the problem by starting with a large network and then reducing it to a smaller network in order to improve generalization. The knowledge of the classifier can be easily extracted in the form of simple classification rules, after its performance is improved by preprocessing and/or pruning. The rule extraction algorithm helps to extract the classification rules from the pruned network as a small trained network is easier to interpret. The thesis has focussed on important approaches which facilitate to build the classifier efficiently using neural networks. It has proposed novel algorithms for discretization, pattern selection and for determining the optimal topology of neural networks. The thesis has also concerned on disclosing the black box nature of neural networks by providing an ideal way to extract the classification rules from the neural classifier.

Firstly, in this research two discretization algorithms such as MDC+PS and DRDS have been proposed and they have been discussed in Chapter 2 and Chapter 3. In MDC+PS method, the proposed discretization method MDC discretizes the continuous attributes into many intervals by the computed mean value but with nominal accuracy. The proposed pattern selection method PS increases the performance of the discretized data obtained using MDC on neural network, in terms of classification accuracy, convergence speed and generalization by obtaining a good training set based on pattern disparity. The results show that the discretization method MDC has to be combined with PS to achieve the best perfor-
mance. To overcome this limitation, a new static, global, incremental, supervised and bottom-up discretization method DRDS has been proposed in chapter 3, based on the scatteredness of data. The measures such as coefficient of dispersion and skewness are used to measure the degree to which numerical data in a range are scattered or stretched. The classification accuracy of the discretized datasets are computed using the feedforward neural network with conjugate gradient training (MLP-CG) algorithm with the help of KEEL software. The results obtained using this discretization algorithm show that the discretization scheme generated by the algorithm almost has minimum number of intervals, requires smallest discretization time and leads to highest classification accuracy.

The quality of the solution found by a neural network depends strongly on the structure of the network used. In this thesis, two pruning methods such as PIHNS and N2PS have been proposed to optimize the architecture of neural network and to improve ANNs classification accuracy and generalization. PIHNS prunes irrelevant hidden neurons by computing the sensitivity using the Euclidian distance. The main advantages of this algorithm are large decrease in number of hidden nodes without affecting the classification accuracy which leads to high degree of generalization and large decrease in computational time for pruning compared with traditional pruning methods. But the main drawbacks of this algorithm are, prunes only the irrelevant hidden neurons and requires user input for pruning parameter. N2PS overcomes these drawbacks by pruning both irrelevant input and hidden neurons based on a significance of a node automatically. The main advantages of this algorithm are, no user defined parameters, large decrease in number
of nodes without affecting the classification accuracy, requires less number of pruning steps and requires less number of iterations for retraining the pruned network compared with other pruning methods and achieves better generalization ability on all datasets. The experimental results demonstrate that the proposed N2PS algorithm is very promising method for determining the optimal architecture of neural networks of arbitrary topology for classifying large datasets.

Lastly, this thesis has also looked at the problem of artificial neural networks, which is their poor interpretability. Rule extraction algorithms translate the internal knowledge of a neural networks into a set of symbolic rules. The proposed work has selected a pedagogical approach for constructing a rule extraction algorithm after considering the complexity of the three approaches namely decompositional, pedagogical and eclectic, which are used for constructing rule extraction algorithms. The pedagogical approach algorithms consist of portability i.e., the rules can be extracted from the network with any structure as it does not analyze the structures/weights of the network. The rule extraction algorithm RxREN, proposed in this research extracts the rules from neural networks using pedagogical approach. The algorithm applies reverse engineering technique to prune the insignificant input neurons and to discover the technological principles of each significant input neuron of neural network used for classification. The results show that RxREN is quite efficient in extracting smaller set of rules with high classification accuracy than those generated by other neural network rule extraction methods. The rules extracted using RxREN are also evaluated by the important measures such as accuracy, fidelity and comprehensibility. The experimental eval-
uations of RxREN indicate that the proposed rule extraction algorithm RxREN is very promising method for discovering the knowledge from neural networks and for interpreting the behaviour of neurons in human understandable format from large datasets with mixed mode attributes of data.

As a summary, this thesis provides novel algorithms that, preprocess the data for classification in data mining, identify the optimal architecture of neural networks for improving classification accuracy and generalization and extract the classification rules of large datasets from trained neural networks. In a nutshell, the algorithms proposed in this research are very effective and easy to use supervised knowledge discovery algorithms which can be applied to problems that require classification of large datasets.

In keeping with this line of research, this section presents some possible avenues for further study. The discretization algorithm proposed in chapter 3 achieves high classification accuracy, but it is not concerned in identifying less number of intervals. This algorithm can be improved by concentrating also on the generation of least number of intervals. In two phases of DRDS, the first phase IDS considers the class information while the second phase FDS does not. Efforts can be taken for generating a final discretization scheme where each interval of discretized attribute potentially has the majority of its values grouped within a single class label, by considering the class attribute interdependency for the merging task of FDS phase and by keeping the stopping criterion as the number of intervals of each discretized attribute to be equal to the number of classes.

The N2PS algorithm, proposed in Chapter 5, determines the optimal structure
of feedforward neural networks by removing irrelevant input and hidden neurons. But still there may exist some irrelevant connections. The formulas of this algorithm can be extended for pruning those irrelevant connections also. N2PS performs well on classification problems with continuous input data. This can be applied also for the binary neural networks with the slight modification. This algorithm requires atmost three retraining steps to obtain the optimal structure. But the methods can be identified for pruning the neural networks without retraining.

The rule extraction algorithm RxREN, proposed in this thesis, extracts the optimal classification rules from the trained neural networks with high accuracy. In future, the proposed method can be assessed in conjunction with other existing pruning algorithms or with applications in which optimized MLP models are available. This thesis has used the iris, ionosphere, heart, diabetes, wave, breastw, credit-g and hepatitis datasets from the machine learning repository to test the proposed algorithms. But the experimental part of this thesis can be extended by testing the proposed knowledge discovery algorithms on many more datasets which are available in this repository and their results can be reported in future research work to establish comparison benchmarks for future research work on discretization, classification and rule mining using these datasets.