Multi-Objective Feature Subset Selection using Non-dominated Sorting Genetic Algorithm

A. Khan*[1] and A. R. Baig[2]

[1]National University of Computers and Emerging Sciences
Islamabad, Pakistan
[2] College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia
*ayeshak1417@yahoo.com

ABSTRACT
This paper presents an evolutionary algorithm based technique to solve multi-objective feature subset selection problem. The data used for classification contains large number of features called attributes. Some of these attributes are not relevant and needs to be eliminated. In classification procedure, each feature has an effect on the accuracy, cost and learning time of the classifier. So, there is a strong requirement to select a subset of the features before building the classifier. This proposed technique treats feature subset selection as multi-objective optimization problem. This research uses one of the latest multi-objective genetic algorithms (NSGA - II). The fitness value of a particular feature subset is measured by using ID3. The testing accuracy acquired is then assigned to the fitness value. This technique is tested on several datasets taken from the UCI machine repository. The experiments demonstrate the feasibility of using NSGA-II for feature subset selection.

Keywords: Optimization, genetic algorithm, classification, Feature subset selection.

1. Introduction

The feature subset selection has become a challenging research area during the past decades, as data sets used for classification purposes in data mining are becoming huge horizontally as well as vertically. Most of the data sets used for classification contain large a number of features (attributes) that are not all relevant. But all these features are used as input to the classification algorithm due to lack of sufficient domain knowledge. Each feature used as a part of the input causes increase in the cost and running time of the classification algorithm and may reduce its generalization ability and accuracy. So there is a huge need for a technique that can find smallest possible feature subset that has high classification accuracy. The multi-objective problems contain more than one objective to be optimized at one time. Most of the real world problems are multi-objective in nature. The feature subset selection problem may also be considered as one of them. The multiple objectives to be optimized simultaneously are the accuracies of the different classes in a data set. Efforts to increase accuracy of one class may reduce the accuracy of another class. This research treats feature subset selection problem as multi-objective problem and uses a multi-objective genetic algorithm to solve it. Multi-class problem has been converted into two-class problem because this research wants to increase the accuracy of each class as a separate objective (equally important). The fitness of each class is evaluated separately, after converting it into two class problem.

There are basically two approaches to solve multi-objective optimization problem. First is ideal multi-objective optimization procedure [14], that finds multiple trade-off optimal solutions and then chooses one of the obtained solutions using higher level information. The second approach is preference-based-multi-objective optimization procedure [14], that first chooses a preference vector and this vector is then used to construct the composite function, which is then optimized to find a single trade-off optimal solution by a single objective optimization algorithm. The most striking difference in single objective and multi-objective optimization is that in multi-objective optimization

the objective functions constitutes an additional multi-dimensional space in addition to the usual decision variable space in the case of single objective function. This additional space is called the objective space [14]. When there are multiple objectives to be optimized simultaneously most researchers make use of the concept of non-dominated or Pareto optimal set of solutions. To understand the concept of non-domination, it is better to understand the concept of domination first. A solution x is said to dominate the other solution y if the solution x is no worse than y in all objectives and the solution x is strictly better than y in at least one objective [14]. There are other approaches using neural network [20], evolutionary multi agent system [21], evolutionary algorithm [22, 25], a hybrid of evolutionary algorithm and neural network [23], approaches using genetic algorithm with particle swarm optimization [24], hybrid evolutionary algorithm [26], and using multi-objective approaches for heuristic optimization [27] for optimization and classification problems. The evolutionary based technique has been used as it works well for the problems with large dimensions, it is known to be a robust technique and it works well with all types of problems because it does not make any assumptions about underlying fitness landscape.

The main features of the proposed method are:

• This research treats feature subset selection as a multi-objective optimization problem.

• The accuracy of each class is considered as a separate objective to be optimized

• This technique makes feature subset selection non-rigid.

• It gives the choice to the user to choose one of the feature subsets in the Pareto-front according to his needs.

• The selected feature subset by the proposed algorithm gives better accuracy and helps to produce less complex classifier.

2. Feature Subset Selection Problem

Feature subset selection is the problem of selecting a subset of features from a larger set of features based on some optimization criteria. Some of the features in the larger set may be irrelevant or mutually redundant. Each feature has an associated measurement cost and risk. So, an irrelevant or redundant feature can increase the cost and risk unnecessarily. The choice of features that represent any data affects several aspects including [15]: *Accuracy:* The features that describing the data must capture the information necessary for the classification. Hence, regardless of the learning algorithm, the amount of information given by the features limits the accuracy of the classification function learned. *Required learning time*: The features describing the data implicitly determine the search space that the learned algorithm must explore. An abundance of irrelevant features can unnecessarily increase the size of the search space and hence the time needed for learning a sufficiently accurate classification function. *Cost*: There is a cost associated with each feature of the data. In medical diagnosis, for example, the data consists of various diagnostic tests. These tests have various costs and risks; for instance, an invasive exploratory surgery can be much more expensive and risky than, say, a blood test. Taking into consideration the above mentioned aspects that are affected by the selection of feature subset, the main objectives for feature subset selection are:

• Improvement in accuracy of the classifier,

• Prediction through a classifier quickly,

• Reduction in the cost

The performance of the classifier depends on many parameters, such as size of training set, number of features and the classifier complexity. If the training set remains the same and number of features increase, the performance of classifier is degraded [9]. As a result, one should minimize the number of irrelevant features which is also known as dimensionality reduction.

3. Multi-objective Evolutionary Algorithm

In feature subset selection problem, multi-objectives comes in naturally. Table 1 shows an example of different feature subsets (solutions) where Di are the features marked as 1 for being present or 0 for being absent. This data has two