Abstract

Feature selection (FS) is a process of selecting a subset of attributes to reduce the data volume without compromising on the performance measures of the data.

The feature selection has become more relevant in the present scenario of data explosion, big data and data mining resulting in reduced computational complexity. The major applications of the feature selection are pre-processing, prediction, data classification, data compression, data analytics and information extraction. There are two basic approaches to the feature selection. The first approach is to determine the redundancy of features among themselves whereas the second approach is to consider the performance of the feature subset for a given objective function. Further, the methods applied for feature selection may follow the following three approaches.

1. Hill climbing
2. Simultaneous selection
3. Heuristics based selection

The present thesis focuses on the simultaneous feature selection approaches for discernibility performance of the data.

In this thesis, three metrics are used to formulate the objective function based on which feature selection is performed. These three metrics (measures) are

(i) classification accuracy based on classification methods such as J48,

(ii) dependency measure based Rough Set (RS), and

(iii) dependency measure based on Lower Approximation based Fuzzy Rough Set (L-FRS).

Rough sets can do classification task in rough sense by capturing the structural relationship within a data. However, this method is applicable to discrete-valued features and

hence continuous valued classification tasks requires discretization of continuous valued features.

Rough set establishes a set of pair of so called upper and lower approximations on a set. The lower approximation is a rough set which classifies the features uniquely i.e. without ambiguity whereas the upper approximation gives a non-unique classification of features. In this work, a POSitive region (POS) defined by lower approximation is used to evaluate a dependency measure which describes the dependency of a given feature set to a class label in a rough sense. This so called rough dependency measure is used as a component in a multi-objective formulation of the problem.

Fuzzy rough set based dependency measure is also used in the objective function due to following motivations.

(i) For the case of real valued features, for applying rough set method, we need to discretize the feature values. The discernibility of features are affected by the quantization and therefore becomes dependent on the quantization of feature values.

(ii) In case of fuzzy rough sets, the real feature values are taken as it is and therefore no such quantization is needed and discernibility of features are therefore more accurate as well as meaningful.

(iii) For real valued features the number of quantization levels can be large or infinite. Thus to capture the feature values in discrete sense is not simply possible.

(iv) If a real valued feature based problem is solved using rough set through discretization, it is highly possible that while application, a newer intermediate value for which the quantization level was not fixed may arise, making the rough set based method of no use since the nominations to values are predefined. In other words, the new value is a new nomination and the rough set based feature reduction has to be done once again including the new nominations.

Chapter 1 of the thesis introduces the problem of feature selection in the context of methods used for the purpose of this thesis. This chapter also introduces methods of Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Intelligent Dynamic Swarm (IDS). The measures of fitness values used in the thesis are also discussed in this chapter.

Some of the swarm intelligence based methods reported in literature for feature selection are particle swarm optimization (PSO), intelligence dynamic swarm (IDS) and ant colony optimization (ACO).

This thesis work presents feature selection methods using swarm based algorithms, with dependency measures as fitness function, using rough and fuzzy rough sets. Inadequacy of hill climbing methods (as they may stuck in local optima) prompted to work with swarm based algorithms and genetic algorithm (GA) for simultaneous selection of features.

Chapter 2 introduces Elitist GA (eGA) based algorithm for feature selection considering classification accuracy as objective function. The performance of eGA on various data sets has been performed and the results have been compared with existing feature selection methods using Fuzzy Rough Set (FRS) based measures. The method of eGA has been applied for feature selection of variables for short-term price forecasting problem.

The fundamentals of Rough Set (FS) and Fuzzy Rough Set (FRS) in the context of feature selection problem has been introduced in Chapter 3. The calculation of performance measures used in the further chapters have been demonstrated in detail.

Chapter 4 discusses the development and implementation of Distributed Sample (DS) initialization proposed in this thesis. This proposed initialization method was implemented in PSO and IDS methods and was tested for its generality.

In Chapter 5 proposes a hybridized version of PSO and IDS. This hybridization of methods has been carried out in the sense of series implementations of PSO and IDS. The implementation is further tested on various data set and its performance has been demonstrated. The advantages of hybridization of PSO and IDS has been demonstrated through comparison with existing PSO and IDS as well as DS initialized PSO and IDS.

Chapter 6 introduces a Butterfly Optimizer (BO) for the purpose of feature selection. As the BO supports real coded variables, two methods are devised to handle requirements of binary coding in the feature selection problem. The results of feature selection on various datasets are compared with the hybrids methods.

Chapter 7 proposes an algorithm named Adaptive Genetic Algorithm with Modified Operator (AGA-MO) for the problem of feature selection. The algorithm is modification of GA in such a way that the mutation operator is adaptively dependent on the distribution of particles in a population. When particles are diverse, the probability of mutation is high and when the population is converging, the mutation is reduced. The test results of application of AGA-MO establishes its superiority over existing eGA and BO algorithms.