Abstract. Considering the importance of the domain relationship in eliminating noisy features in feature selection, we present an alternate approach to designing a multi-objective fitness function using multiple correlation for the genetic algorithm (GA), which is used as a search tool in the problem. Multiple correlation is a simple statistical technique that uses the multiple correlation coefficients to measure the relationship between a dependent variable and a set of independent variables within the domain space. Simulation studies were conducted on both real-world and controlled data sets to assess the performance of the proposed fitness function. The comparison between the traditional fitness function and our proposed function is also reported. The results show that the proposed fitness function can perform more satisfactorily than the traditional one in all cases considered, including different data types, multi-class and multi-dimensional data.

Keywords: Feature selection, genetic algorithm, fitness function, domain relationship, multiple correlation.

1 Introduction

Feature selection is a very important part of classification system preprocessing. The choices of the feature subset affect classification accuracy, training time and the cost of data collection. Absence of necessary information limits classification accuracy, while irrelevant and/or redundant features increase training time, including searching and learning processes, due to the large size of the search
space. More features also cost more in data collection procedures. The aim of feature selection is to reduce the number of irrelevant and/or redundant features as much as possible without losing performance. The fewer the number of features, the simpler the system, and therefore faster the computation. This leads to overall performance improvement. However, in real-world situations, problems are very complex. In feature selection problems, relevant features are unknown \textit{a priori}. In reality, additional candidate features are introduced to improve domain representation, but many of them are irrelevant and/or redundant features. Moreover, in applications, the domain space can be extremely large. Even for a medium size feature set, there are $2^D$ possible subsets for any $D$ number of features. For these reasons, the need for an efficient search method and learning algorithm is of much importance.

Many methods have been proposed for feature subset selection. Almuallim and Dietterich [2, 3] introduced FOCUS-2 to implement the MIN-FEATURES bias and the Mutual-Information-Greedy, Simple-Greedy and Weight-Greedy algorithms to apply efficient heuristics for approximating the MIN-FEATURES bias. Kira and Rendell [15, 16] used Relief algorithms to efficiently approximate the set of relevant features. Caruana and Freitag [5] presented a caching scheme that makes feature hill climbing a more practical computation. Skalak [31] replaced greedy search with random hill climbing. Liu and Setiono [18–23] have proposed different algorithms for feature selection. A taxonomy of available feature selection algorithms into broad categories is presented in [13]. This study focuses on the multiple solutions method in the taxonomy. The method has been divided into two submethods: deterministic (e.g. beam search) and stochastic (e.g. genetic algorithm (GA)) [13]. Siedlecki and Sklansky [29, 30] have discussed beam search and introduced GA for feature selection. Vafaie and De Jong [33] have suggested that GA can be used to increase the robustness of feature selection algorithms because of its ability to efficiently search a large space and population using probabilistic selection, which is difficult or even impossible to do with domain knowledge and theory. Choosing an appropriate fitness function is an essential step for a successful GA application in any problem domain [34]. In a feature selection, predictions can reach minimal accuracy if the necessary features are not included in the feature subset. This could be caused by lack of information. To avoid this unexpected case in real applications, an adequate fitness function covering multi-objectives should be more general and able to guide in searching for the best feature subset for the feature selection problem domain. The fitness function must be able to measure and maximize the performance of the feature subset in discriminating classes of pattern space, while minimizing their sizes.

There are many related works on GA in feature selection. Punch \textit{et al.} [26] summarized works that combine feature selection and data classification using genetic algorithms. Vafaie and Imam [36] have presented a comparison between the Importance Score (IS), which is based on a greedy-like search, and a genetic algorithm-based method. They addressed an interesting open question regarding whether a multistrategy approach could be developed that can combine the two