Information Processing Letters 117 (2017) 25–34

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

A novel adaptive feature selector for supervised classification

S. Sasikala [a,*], S. Appavu alias Balamurugan [a], S. Geetha [b]

[a] K.L.N. College of Information Technology, Tamil Nadu, India
[b] School of Computing Science and Engg., VIT University, Chennai Campus, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 26 April 2014
Received in revised form 19 May 2016
Accepted 2 August 2016
Available online 8 August 2016
Communicated by W.-L. Hsu

Keywords:
Medical data mining
Biomedical classification
Feature selection
Genetic algorithm
Performance evaluation

ABSTRACT

Optimal feature Selection is an imperative area of research in medical data mining systems. Feature selection is an important factor that boosts-up the classification accuracy. In this paper we have proposed a adaptive feature selector based on game theory and optimization approach for an investigation on the improvement of the detection accuracy and optimal feature subset selection. Particularly, the embedded Shapely Value includes two memetic operators namely include and remove features (or genes) to realize the genetic algorithm (GA) solution. The use of GA for feature selection facilitates quick improvement in the solution through a fine tune search. An extensive experimental comparison on 22 benchmark datasets (both synthetic and microarray) from UCI Machine Learning repository and Kent ridge repository of proposed method and other conventional learning methods such as Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), J48 (C4.5) and Artificial Neural Network (ANN) confirms that the proposed SVEGA strategy is effective and efficient in removing irrelevant and redundant features. We provide representative methods from each of wrapper, filter, conventional GA and show that this novel memetic algorithm – SVEGA yields overall promising results in terms of the evaluation criteria namely classification accuracy, number of selected genes, running time and other metrics over conventional feature selection methods.

1. Introduction and literature review

Medical data mining has revealed to be a flourishing tactic in medical field as an aid for the clinical and diagnosis data [1]. Currently utilizing data mining in health care analysis is acquiring more importance. The easy accessibility of medical data for analysis and diagnosis of disease provides medical data mining as a dais for health care of the patient.

Feature selection methods [2,5,6] have a propensity to spot the prominent features those are more significant for classification and they can be generally classified as either feature subset selection methods or feature ranking methods. The Feature Subset Selection (FSS) methods return a subset of the original set of features which are regard as to be the most significant for classification. Feature Ranking (FR) methods ranks the features according to their usefulness in the classification task. On comparing with feature subset selection methods, feature ranking methods were widely used in many application domains. In dimensionality reduction, feature selection is used as pre-processing task to machine learning to remove irrelevant and redundant data for increasing classifier accuracy. The intention of this research work is aimed at showing that selection of more important features from the available unprocessed medical dataset which helps the medical doctor to turn up at an accurate diagnosis. The chief focus is on aggressive dimensionality reduction so as to find our self with an increase in the prediction accuracy. The features undergo a

* Corresponding author.
E-mail addresses: nithilannsasikala@yahoo.co.in (S. Sasikala), app_s@yahoo.com (S. Appavu alias Balamurugan), geethabaalan@gmail.com (S. Geetha).

http://dx.doi.org/10.1016/j.ipl.2016.08.003
0020-0190/

genetic evolution process along with memetic operations to select the features which forms a feature subset with lowest cardinality.

The objective of this research work is aimed at showing that selection of more significant features from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The primary focus is on aggressive dimensionality reduction so as to land-up with increase in the prediction accuracy. The features are subjected to a genetic evolution process within which they undergo the memetic operations namely include and *remove*, at the end of which, only the features that increase the accuracy, and form the subset with the lowest cardinality, are obtained. The empirical results show that the proposed SVEGA feature selector achieves remarkable dimensionality reduction in the 22 medical (synthetic and microarray) datasets obtained from the UCI Machine Learning repository [3] and Kent ridge repository [4].

After reviewing the works on feature selection for medical dataset [13] it is observed that most of the existing methods suffer from the following problems: (1) depending on the complexity of the search method, the iterations of evaluations are too large; (2) they rely on a univariate ranking that does not take into account interaction between the variables already included in the selected subset and the remaining ones. Moreover, a method that produces the best accuracy employs more number of features and hence more running time is involved in the construction of the respective classifiers. Contrarily, a method that outputs the fewest number of features produces inferior detection accuracy. A holistic and universal method that achieves the best classification accuracy with fewest features possible is still an open research problem. This paper makes an attempt to design such a feature selection sequence and it is called as "Shapely Value Embedded Genetic Algorithm (SVEGA)" and its performance is compared with already existing methods for feature selection. Their work is reviewed for the clear explanation. Bu et al. propose a hybrid feature selection algorithm which combines ReliefF algorithm with Shapley value analysis for selecting highest relevant features based on estimated Shapley value. Tan et al. apply Genetic Algorithm (GA) for feature subset selection based on multiple feature selection criteria and find small subsets of features that perform well for the inductive learning algorithm for building the classifier with classification accuracy. As another competitor memetically framed novel hybrid feature selection algorithm (MA-C) for feature selection is presented by Senthamarai Kannan and Ramaraj [17]. The selected features maximized the effectiveness and efficiency of the proposed hybrid filter and wrapper feature selection algorithm for classification problem using a memetic framework.

2. System and methodology

Shapley Value Analysis has been proved to be a promising strategy for feature selection process. Shapley Value Analysis (SVA) [12,14] is a game theory based technique for causal function localization that addresses the issue in describing and calculating the contributions made by the interactions among the group of elements in a data set with multiple features and their corresponding performance scores. In this section, some basic concepts related to game theory and shapely values are touched upon. Later their relevance to feature selection is introduced. In game theory, a **cooperative game** is a game where groups of players ("coalitions") may enforce cooperative behavior and aim to obtain high total profit.

Shapley is the value of an operator that assigns an expected marginal contribution to each player in the game with respect to a uniform distribution over the set of all permutations on the set of players. Specifically, let Π be a permutation (or an order) on the set of players, i.e., a mapping exists as one-to-one function from N onto N, and let us imagine the players appearing one by one to collect their payoff according to the order Π. The marginal contribution Δ_i of player i to a coalition S is given as follows:

$$\Delta_i(s) = v\big(s \cup \{i\}\big) - v(s) \tag{1}$$

Here function v associates with every non-empty subset S of F, a real number $v(s)$ (the value of S) with $v(\{\varphi\}) = 0$. The unbiased estimator for the Shapley value, for a player i is given by the mean of marginal contributions to all possible coalitions of players in N, is given as

$$\Phi_i(v) = \frac{1}{n!} \sum_{x \in \Pi} \Delta_i\big(S_i(\pi)\big) \tag{2}$$

where Π is the set of permutations over N and $S_i(\pi)$ is the set of players from π that appear before player i in the permutation.

Generally Shapley value is widely analyzed concept under the game theory. Here these general axioms of Shapley value are stated to specify the analogy (comparison) of ways in performing the feature selection by eliminating the irrelevant (unwanted or null), redundant (duplicate) features and also performing the sorting of the features.

2.1. Shapley value – embedded GA

In this section, the proposed memetic algorithm, particularly, Shapley Value Embedded GA (SVEGA) is outlined. At the beginning of the SVEGA search, the population for GA [16] is initialized randomly where each chromosome in the pool encodes a candidate feature subset. In this work, each chromosome is built of a binary string whose length equals the total number of features in the dataset of interest. In binary encoding, a bit of value '1' ('0') indicates that the respective feature is selected (omitted). The objective function for calculating the fitness of each chromosome is then obtained as follows:

$$Fitness(c) = \mathrm{Max}\big(Obj\_Fun(SF_c)\big) \tag{3}$$

$$\begin{aligned} Obj\_Fun(SF_c) &= \alpha * (1/\tau) + (\mathrm{RCF} * \mathrm{Recall}) \\ &\quad + (\mathrm{PCF} * \mathrm{Precision}) \end{aligned}$$

$$\begin{aligned} &\text{where } \tau = \text{No. of ones in the } SF_c \\ &\alpha = \text{No. of minimum features selected} \\ &\mathrm{RCF} = \text{Recall Credibility Factor} \\ &\mathrm{PCF} = \text{Precision Credibility Factor} \end{aligned} \tag{4}$$