Knowledge-Based Systems 22 (2009) 356–362

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A filter model for feature subset selection based on genetic algorithm

M.E. ElAlami

Department of Computer Science, Mansoura University, Mansoura 35111, Egypt

ARTICLE INFO

Article history:
Received 15 January 2008
Received in revised form 27 September 2008
Accepted 17 February 2009
Available online 26 February 2009

Keywords:
Feature subset selection
Relevant feature
Genetic algorithm
Artificial neural networks
Non-linear optimization
Fitness function

ABSTRACT

This paper describes a novel feature subset selection algorithm, which utilizes a genetic algorithm (GA) to optimize the output nodes of trained artificial neural network (ANN). The new algorithm does not depend on the ANN training algorithms or modify the training results. The two groups of weights between input-hidden and hidden-output layers are extracted after training the ANN on a given database. The general formula for each output node (class) of ANN is then generated. This formula depends only on input features because the two groups of weights are constant. This dependency is represented by a non-linear exponential function. The GA is involved to find the optimal relevant features, which maximize the output function for each class. The dominant features in all classes are the features subset to be selected from the input feature group.

1. Introduction

Reducing dimensionality of a problem, in many real world problems, is an essential step before any analysis of the data. The general criterion for reducing the dimensionality is the desire to preserve most of the relevant information of the original data according to some optimality criteria [1]. Dimensionality reduction or feature selection has been an active research area in pattern recognition, statistics and data mining communities. The main idea of feature selection is to choose a subset of input features by eliminating features with little or no predictive information. In particular, feature selection removes irrelevant features, increases efficiency of learning tasks, improves learning performance and enhances comprehensibility of learned results [2,3]. Feature selection problem can be viewed as a special case of the feature-weighting problem. The weight associated with a feature measures its relevance or significance in the classification task [4]. If we restrict the weights to be binary valued, the feature-weighting problem reduces to the feature selection problem. Feature selection algorithms fall into two broad categories, the filter model or the wrapper model [5]. Filter models use an evaluation function that relies solely on properties of the data, thus it is independent on any particular algorithm. Wrapper models use the inductive algorithm to estimate the value of a given subset. Most algorithms for feature selection perform either heuristic or exhaustive search [6]. Heuristic feature selection algorithms estimate the feature's quality with a heuristic measure such as information gain [7], Gini index [8], discrepancies measure [9] and chi-square test [10]. Other examples of heuristic algorithms include the Relief algorithm and its extension [11]. Exhaustive feature selection algorithms search all possible combinations of features and aim at finding a minimal combination of features that are sufficient to construct a model consistent with a given set of instances such as the FOCUS algorithm [12]. Various approaches have been proposed for finding irrelevant features and remove them from the feature set. C4.5 decision tree presented in [13] finds relevant features by keeping only those features that appear in the decision tree. The cross-validation method is applied in [14] to filter irrelevant features before constructing ID3 and C4.5 decision trees. Neural network is used to estimate the relative importance of each feature (with respect to the classification task), and assign it a corresponding weight [15]. When properly weighted, an important feature would receive a larger weight than less important or irrelevant features. In general, feature selection refers to the study of algorithms that select an optimal subset from the input feature set. Optimality is normally dependent on the evaluation criteria or the application's needs. Therefore, the genetic algorithms (GAs) have recently received much attention because of their ability to solve difficult problems in the optimization. The GAs are search methods that have been widely used in feature selection where the size of the search space is large [16]. The most important differences between the GAs and the traditional optimization algorithms are: genetic algorithms work with a coded version of the parameters; they do not search from one single point, but from a population of points. A crucial issue in the design of a genetic algorithm is the choice of the fitness function. This function is used to evaluate the quality of each

E-mail address: Moh_ElAlmi@mans.eun.eg

0950-7051/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2009.02.006

hypothesis, and it is the function to be optimized in the target problem.

This paper presents a novel algorithm for feature subset selection from trained neural network using genetic algorithm. It does not depend on the ANN training algorithms and it does not modify the training results. The GA is used to find the optimal input features (relevant), which maximize the output functions of trained neural network.

The organization of this paper is as follows. The problem formulation is described in Section 2. The data preprocessing is performed in Section 3. The proposed feature selection algorithm is outlined in Section 4. An initial experiment is described in Section 5 to demonstrate the feasibility of the proposed algorithm. The application and results are reported in Section 6. The conclusion and future work are presented in Section 7.

2. Problem description

The proposed feature subset selection algorithm starts with training the artificial neural network on the input features and the corresponding class. The ANN is trained so that a satisfactory error level is reached. Each input unit corresponds typically to a single feature and each output unit corresponds to a class value. The main objective of our approach is to encode the network in such a way that a genetic algorithm can run over it. After training the ANN, the weights between input-hidden and hidden-output layers are extracted. Therefore, each output node of ANN can be represented as a general function of input features and extracted weights. The activation function used in the hidden and output nodes of the ANN is a sigmoid function. Therefore, each output function is non-linear exponential function, which has a maximum output value of one. For each output node, the GA is used to find the optimal values of the input features, which maximize the output function. The obtained features represent the relevant features for each class value. The dominant features in all classes are the overall relevant features for a given database. The proposed algorithm for feature selection is shown in Fig. 1.

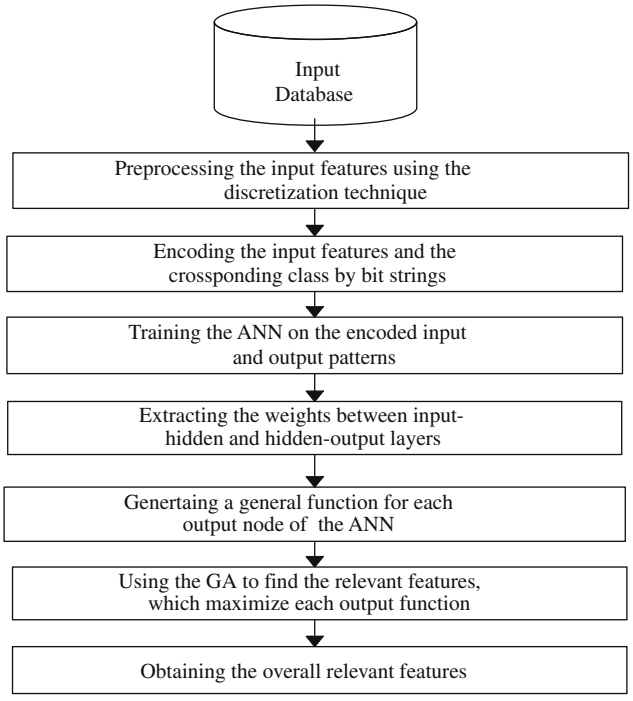

Fig. 1. The framework of the proposed feature subset selection algorithm.

3. Data preprocessing

Some of the learning algorithms such as neural network are often trained more successful and fast when the discrete input features are used. Therefore, the features which have numerical values in a given database must be treated by using the discretization technique. The discretization technique splits the continuous feature values into small sets of intervals, each interval has upper and lower values $[X_{lower}–X_{upper}]$. Hence, these intervals are transformed into linguistic terms. The discretization of features is formulated as an optimization problem. The GA is used for obtaining the optimal boundaries of these intervals which maximize the density of the predominate class while minimizing the other classes densities in the given interval. As a result, all features in the given database can be transformed into discrete values via substituting each interval by a linguistic term such as short (S), medium (M) or large (L). The mathematical model of this technique and its applications is presented in my paper [17].

4. The proposed algorithm

A supervised ANN uses a set of M examples or records. These records include N features. Each feature, F_n (n = 1,2,...,N), can be encoded into a fixed length binary sub-string $\{x_1\ x_i\ x_{Vn}\}$, where V_n is the number of possible values of nth feature. The element x_i = 1 if its corresponding feature value exists, while all the other elements = 0. Therefore, the proposed number of input nodes, I, in the input layer of ANN is given as:

$$I = \sum_{n=1}^{N} V_n \tag{1}$$

Consequently, the input features vector, X_m, to the input layer can be rewritten as:

$$X_m = \{x_1 \quad x_i \quad x_I\} \tag{2}$$

The output class vector, T_m is encoded as a bit vector of a fixed length K as follows:

$$T_m = \{O_1 \quad O_k \quad O_K\} \tag{3}$$

where m = (1,2,...,M), M is the total number of input training patterns; k = (1,2,...,K), K is the number of different possible classes.

If the output vector belongs to class$_k$ then the element O_k is equal to one while all the other elements in the vector are zeros. Therefore, the proposed number of output nodes in the output layer of ANN is K. Accordingly, the input and output nodes of the ANN are determined and the structure of the ANN is shown in

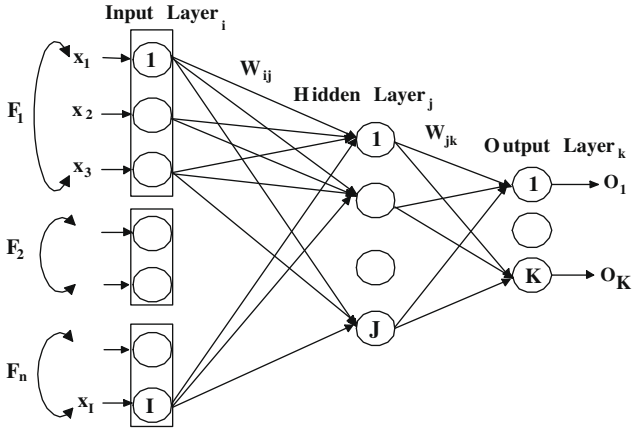

Fig. 2. The structure of the artificial neural network.