Biomedical Research 2016; Special Issue: S99-S102

ISSN 0970-938X
www.biomedres.info

Multi-ranked feature selection algorithm for effective breast cancer detection.

R Jaya Suji[1]*, SP Rajagopalan[2]

[1]Sathyabama University, Chennai, India

[2]GKM College of Engineering and Technology, Chennai, India

Abstract

In this paper, we propose a new feature selection algorithm called Multi-Ranked Feature Selection Algorithm (MRFSA) for effective feature selection to be used in a new cancer detection system. In addition, we use a Neuro-Fuzzy Temporal Classification algorithm for testing the performance of the feature selection algorithm in terms of classification accuracy. To test the efficacy of the proposed medical expert system, we have used the Wisconsin Breast Cancer Dataset (WBCD). The obtained classification accuracy is a very promising result compared to the existing works in this area based on the reports from the results for the same data set.

Keywords: Neuro-fuzzy, Classification, Feature selection, Ranking, Nearest neighbor.

Accepted on August 03, 2016

Introduction

Breast cancer is one of the most commonly occurring cancers in women. Early diagnosis of breast cancer is crucial and important in reducing mortality rate and improving the patient's quality of life. An early diagnosis of this disease has more importance and considerably improves the prognosis and leads to more effective treatment for the patient. Expert system is a computer system that uses artificial intelligence to emulate the decision-making ability of a human expert and to solve problems within a specialized domain that ordinarily requires human expertise. In order to achieve the feats of apparent intelligence, an expert system relies on two components called knowledge base and an inference engine. A knowledge base is an organized as a collection of facts about the system's domain. An inference engine interprets and evaluates the facts in the knowledge base in order to provide an effective solution. Typical tasks for expert systems involve classification, diagnosis, monitoring, and design, scheduling, and planning for specialized endeavors.

Literature Survey

Many works have been carried out in the literature for effective feature selection and the detection of breast cancer [1-4]. Among them, Ganapathy et al., proposed new intelligent agent based attribute selection algorithm for effective feature selection and also proposed an effective classification algorithm called Intelligent Enhanced Multiclass Support Vector Machine (EMSVM) [3]. They achieved better classification accuracy by using the selection of optimal number of features. The same authors also enhanced these two algorithms in better manner in 2013 in terms of detection accuracy. El-Khatib proposed new Feature Reduction algorithm for better detection accuracy in their decision making system [1]. Ghaemi et al. introduced a new optimization algorithm called Forest Optimization Algorithm for optimizing the number of features in a standard dataset which is useful for improving the classification accuracy [2].

Di-Leo et al. discussed about new approaches for improving the outcomes in breast cancer in Europe in terms of detection accuracy [5]. They also discussed about the precautions to prevent the breast cancer disease. Mani et al. applied an objective and reproducible immunofluorescence-based assay to quantify the distribution of Tumor-infiltrating lymphocyte (TIL) expression in multiple spatially separate regions from a population of breast tumors [6]. Buchsbaum et al. explain in detail about the breast cancer is a death disease and also must be taken remedial actions against the breast cancer [7]. They also discussed some necessary steps for preventing the breast cancer disease. However, the existing systems are not able to identify the optimal number of features leading to increase in complexity of the classifier. Hence, a new feature selection algorithm is necessary to identify optimal number of features. In this paper, we propose a new Multi-Ranked feature selection algorithm for effective feature selection to achieve the better classification accuracy over the medical dataset. The major contribution of this paper is the introduction of multi-ranking methodology for ranking the features efficiently. Due to efficient feature selection, the proposed system achieves better classification accuracy.

Proposed Work

In this section, we discusses about the proposed feature selection algorithm called multi-ranked feature selection algorithm and Neuro-Fuzzy Temporal classifier.

Feature Selection

The proposed weighted subset feature selection algorithm is proposed according to [1] with the help of efficient optimization technique called FOREST optimization algorithm [2] and effective feature subset based classifier called Enhanced Multiclass Support Vector Machine [3] algorithm for validating the classification accuracy of selected feature subset.

In this proposed feature selection algorithm, first calculate the Information Gain Ratio (IGR) value for all features by using the equations (1-3) according to Ganapathy et al. [3].

$$Info\,(D) = -\left[\frac{[freq(C_j, D)}{|D|}\right]\log_2\left[\frac{[freq(C_j, D)}{|D|}\right] \rightarrow (1)$$

$$Info\,(T) = \left[\frac{|T_i|}{|T|}\right] \times Info(T_i) \rightarrow (2)$$

$$IGR(A_i) = \left[\frac{Info(D) - Info(T)}{Info(D) + Info(T)}\right] \times 100 \rightarrow (3)$$

Where D indicates the total dataset, C_j indicates the particular feature of the cluster (Subset), T indicates the total training dataset considered for IGR calculation, T_i indicates the features of the training dataset. And Ai is indicates the attribute (feature) of the particular record.

Forest optimization technique

The existing optimization algorithm called Forest Optimization Algorithm (FOA) [2] is used in this work for re-ranking the features. Here, we have calculated the feature weights according to the age calculation of FOA. In this way, we have considered each patient record of the dataset as a tree. These features are ranked using the local seeding, global seeding and fitness function. Finally, this algorithm has obtained the best subset of the given medical dataset according to the re-ranking of the features.

Enhanced multiclass support vector machine

We have used the Enhanced Multiclass Support Vector Machine (EMSVM) [3], for evaluating the selected feature subsets during the process of feature selection. In this work, this classification algorithm helps to calculate the detection accuracy based on many aspects of the problem. It is capable of solving the multiclass problem efficiently.

The multi-ranked feature selection algorithm (MRFSA)

In this paper, we propose a new feature selection algorithm called Multi-Ranked Feature Selection Algorithm (MRFSA) with the help of IGR, Forest Optimization Algorithm [2] and EMSVM [3]. Here, the proposed algorithm chooses a set of optimal features as input to the forest optimization algorithm for selecting the valuable features from the full set of features. These selected features are evaluated by using EMSVM. If the selected feature subset provides better classification accuracy then the process is stopped otherwise it is repeated until it provides classification accuracy above a prescribed level.

The proposed feature selection algorithm works as follows:

Input: F-Full set of features, IGR: Information Gain Ratio, C: Classifier, T: Threshold

Output: Optimal features

Algorithm:

Step 1: Read all the features F_i

Step 2: Calculate the Information Gain Ratio value for all features using the equations (1-3).

Step 3: Rank the features according to the IGR value of the feature.

Step 4: Initialize an empty set S = {}, ac=0;

Step 5: Initialize the value of ac in to ap.

Step 6: Read the next feature and assign into a temporary variable f.

Step 7: Append this feature into the new feature Subset.

Step 8: Remove the feature from original feature set.

Step 9: Find the accuracy for the new feature subset using EMSVM [3] and store into ac.

Step 10: If ((ac-ap) < T OR ac <ap) then Go to Step 10.

Step 11: Otherwise re-rank the selected feature subset using FOA [2]

Step 12: Repeat step 5 to Step 9

Step 13: List the optimal features.

The proposed algorithm helps to identify the optimal features which are helpful for making effective decisions over the medical dataset. The proposed algorithm provides better prediction accuracy over the medical dataset due to the uses of multi-ranked features.

Classification

In this work, classification is performed using the Neuro-Fuzzy Temporal classification algorithm [4] on the selected features. Moreover, this algorithm has been tested using WDBC dataset. From the combination of the proposed feature selection algorithm and the classifier used in this work, it is observed that the proposed feature selection algorithm when it is used with the classifier has reduced the classification time and increased the classification accuracy. The classification algorithms are used to categorize the data. The rules are used to finalize the category of the data. Some of the classification algorithms such as Multiclass Support Vector Machine used distance measurement formula to find the distance between the data value and consider this value for making initial decisions. After taken the initial decision, the final decision will be taken by using the rules which are framed according to the problem by the algorithm developer. In this paper, we have used MSVM for making effective decision over the WDBC dataset. For improving the classification accuracy of the existing MSVM classification algorithm, we have proposed new feature selection algorithm called Multi-ranked feature selection algorithm. The proposed feature selection algorithm helps to improve the classification accuracy by using optimal number of features. The proposed feature selection algorithm is also reduces the classification time.

We have used the Wisconsin Breast Cancer Dataset (WBCD) taken from the UCI Machine Learning repository in order to evaluate the effectiveness of the proposed classification model [8,9]. We have evaluated the proposed feature selection algorithm by using this dataset. Here, we have used three evaluation metrics such as specificity, sensitivity and f-measure for finding the classification accuracy. The experiments have been conducted by using JAVA programming language. We have considered the WBCD dataset for evaluating the proposed algorithm. The standard dataset is divided into two sets (60% and 40%), one for training and another one set for testing. Five experiments have been conducted for evaluating the proposed system with different size of datasets with full features. Figure 1 shows the performance analysis of the proposed system with full features and selected features.

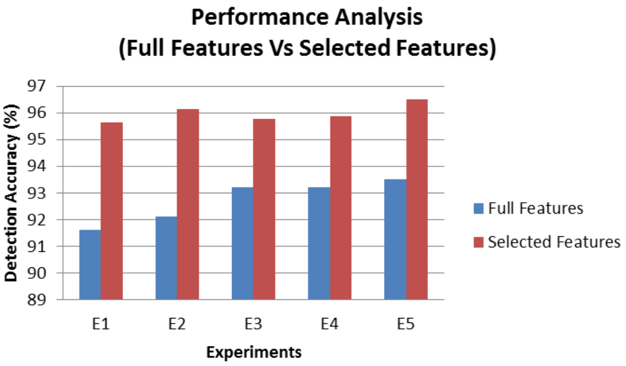

Figure 1. Performance analysis (full features vs. selected features).

From Figure 1, it can be observed that the performance of the proposed system with selected features is better when it is compared with the proposed system with full features. This is due to the fact that the use of most important features. Table 1 shows the performance comparative analysis of the proposed feature selection algorithm and the existing feature selection algorithms. The existing algorithms have been tested with WDBC dataset.

Table 1. Performance compative analysis.

S.No .	Feature Selection Techniques	No. of Features Selected	Classification Accuracy	
			Full Features	Selected Features
1	IAASA [3]	11	88.4	91.4
2	PSO [4]	9	90.7	93.5
3	Genetic Algorithm [10]	8	91.8	94.7
4	Proposed MRFSA	7	92.72	95.98

From Table 1, it can be seen that the proposed feature selection algorithm is performed well than the existing feature selection algorithms. This is due to the fact that the use of ranking algorithm and the MSVM algorithm. Here, the classification process is helpful for selecting optimal number of features. Figure 2 shows the comparative analysis between the proposed system and the existing systems which are developed by various researchers in the past in this direction. The standard and famous classification and optimization algorithms are considered for comparative analysis to know the efficiency of the proposed system.

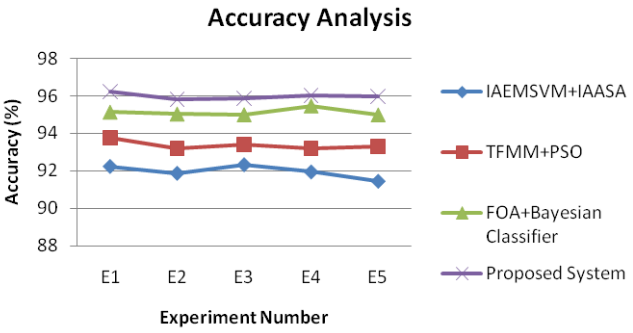

Figure 2. Accuracy analysis.

From Figure 2, it can be observed that the proposed classifier provides more accuracy than the existing classification algorithms due to the fact that it uses the multi-ranked features. The combination of multi-ranking and multiclass classification is also used to improve the overall accuracy of the proposed system.

Conclusion and Future Enhancement

In this paper, a new Multi-Ranked feature selection algorithm for effective feature selection to achieve the better classification accuracy over the medical dataset is proposed. By introducing the new feature selection algorithm, different classification algorithms were tested to find the classification accuracy. The introduction of multi-ranked feature selection is very useful to identify the most important features for better classification accuracy. Finally, the best classifier for this application was identified as Neuro-Fuzzy Temporal Classification algorithm which is applied in this work for identifying breast cancer. The major contribution of this paper is the introduction of the new algorithm for feature ranking which performs the ranking of features efficiently for achieving better classification accuracy. The classification accuracy difference between the proposed system and the existing systems is more than 1%. This is reasonable improvement in the field of breast cancer disease detection. Future works in this direction can be the introduction of intelligent agents to make a distributed system in which the proposed feature selection algorithm can be used.

References

1. El-Khatib K. Impact of Feature Reduction on the Efficiency of Wireless Intrusion Detection Systems. IEEE Transact Parallel Distribut Syst 2010; 21: 1143-1149.
2. Ghaemi M, Feizi-Derakhshi MR. Forest Optimization Algorithm. Exp Syst Appl 2014; 41: 6676-6687.
3. Ganapathy S, Yogesh P, Kannan A. An Intelligent Intrusion Detection System for Mobile Ad-Hoc Networks Using Classification Techniques. Comput Commun Informa Syst 2011; 148: 117-122.
4. Ganapathy S, Sethukkarasi R, Yogesh R, Vijayakumar P, Kannan A. An intelligent temporal pattern classification system usingfuzzy temporal rules and particle swarm optimization. Sadhana 2014; 39: 283-302.
5. Di Leo A, Curigliano G, Dieras V, Malorni L, Sotiriou C, Swanton C, Thompson A, Tutt A, Piccart M. New approaches for improving outcomes in breast cancer in Europe. Breast 2015; 24: 321-330.
6. Mani NL, Schalper KA, Hatzis C, Saglam O, Tavassoli F, Butler M, Chagpar AB, Pusztai L, Rimm DL. Quantitative assessment of the spatial heterogeneity of tumor-infiltrating lymphocytes in breast cancer. Breast Cancer Res 2016; 18: 1-10.
7. Buchsbaum RJ, Oh SY. Breast Cancer-Associated Fibroblasts: Where We Are and Where We Need to Go. Cancers 2016; 8: 1-19.
8. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, San Jose, CA, 1993; 1905: 861-870.
9. Mangasarian OL, Street WN, Wolberg WH. Breast cancer diagnosis and prognosis via linear programming. Operat Res 1995; 43: 570-577.
10. Sindhu SS, Geetha S, Kannan A. Decision tree based light weight intrusion detection using a wrapper approach. Expert Syst Appl 2012; 39: 129-141.

*Correspondence to

R Jaya Suji

Sathyabama University

India