Digital Signal Processing 23 (2013) 230–237

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Genetic algorithm wrapped Bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases

Akın Özçift [a,*], Arif Gülten [b]

[a] Computer Programming Division, Vocational High School, University of Gaziantep, Turkey
[b] Electrical and Electronics Engineering Department, Fırat University, Elazığ, Turkey

ARTICLE INFO

Article history:
Available online 23 July 2012

Keywords:
Erythemato-squamous
Genetic algorithm
Wrapper feature selection
Bayesian network
Best first search
Sequential floating search
Medical diagnosis

ABSTRACT

This paper presents a new method for differential diagnosis of erythemato-squamous diseases based on Genetic Algorithm (GA) wrapped Bayesian Network (BN) Feature Selection (FS). With this aim, a GA based FS algorithm combined in parallel with a BN classifier is proposed.
Basically, erythemato-squamous dataset contains six dermatological diseases defined with 34 features. In GA–BN algorithm, GA makes a heuristic search to find most relevant feature model that increase accuracy of BN algorithm with the use of a 10-fold cross-validation strategy. The subsets of features are sequentially used to identify six dermatological diseases via a BN fitting the corresponding data. The algorithm, in this case, produces 99.20% classification accuracy in the diagnosis of erythemato-squamous diseases. The strength of feature model generated for BN is furthermore tested with the use of Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Simple Logistics (SL) and Functional Decision Tree (FT). The resultant classification accuracies of algorithms are 98.36%, 97.00%, 98.36% and 97.81% respectively. On the other hand, BN algorithm with classification accuracy of 99.20% is quite a high diagnosis performance for erythemato-squamous diseases. The proposed algorithm makes no more than 3 misclassifications out of 366 instances. Furthermore, FS power of GA is also compared with two alternative search algorithms, i.e. Best First (BF) and Sequential Floating (SF).
The obtained results have all together shown that the proposed GA–BN based FS and prediction strategy is very promising in diagnosis of erythemato-squamous diseases.

1. Introduction

Automated medical diagnosis strategies mainly rely on a Machine Learning (ML) algorithm that is trained to learn former decision characteristics of an physician about a specific disease and then it can be used to assist human-decision makers to diagnose future patients of the same disease [1,4]. Unfortunately, there is no universal ML model that can adapt itself to any kind of disease. In most cases, to develop a novel automated diagnosis system, developers uses a three-step design strategy: (i) select a specific disease dataset and prepare it no next step, (ii) if data is high dimensional, apply feature reduction strategies in accordance with the third step and finally (iii) investigate an ML strategy with highest possible accuracy [4].

High number of features in a dataset can lead to lower classification accuracy with high computational cost and risk of "over-fitting" [5,11]. In other words, smaller number of features might increase classification accuracy with decreased computational cost and it eliminates risk of "over-fitting". Therefore, elimination of irrelevant features from a high dimensional dataset is a fundamental step for designing automated diagnosis systems with high-accuracy. Hence, the main objective of this study is to design a FS method to reduce dimension of erythemato-squamous diseases dataset and to obtain high-accurate classification rates. We therefore developed a two step FS and classification algorithm: In the first step, the dimension of data is reduced with the FS policy, i.e. GA algorithm running in parallel with BN. Consequently, once the new feature model is obtained, a BN classifier is built for this reduced dataset to measure feature model strength. A stratified 10-fold cross-validation strategy is used while the obtained feature model is validated. The strength of the feature subset is also tested with the use of four additional ML algorithms, i.e. SVM, MLP, SL and FT. Furthermore, the proposed algorithm is compared with two alternative wrapper search algorithms namely BF and SF.

The differential diagnosis of erythemato-squamous diseases is relatively a difficult problem of dermatology, since this group of diseases, i.e. psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris, all have the clinical features of erythema and scaling with minor differences. For diagnosis, a biopsy is in general necessary; however this group of diseases shares also numerous histopathological features

* Corresponding author.
E-mail address: akinozcift@hotmail.com (A. Özçift).

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2012.07.008

as well. Furthermore, at the beginning phase, the disease might exhibit the features of another disease and it may have the particular characteristics in advance. The dermatological patients were first consulted clinically with 12 features. For analysis of remaining 22 histopathological features, skin samples were taken from patients. These histopathological features are examined by the use of a microscope [2,3].

The rest of the paper is organized as follows. The next section presents a literature survey for erythemato-squamous disease problem. Section 3 provides information about the dermatology disease and it is followed by a brief introduction to FS strategies. ML methods to evaluate strength of the feature model obtained by GA–BN algorithm are explained in Section 5. Model validation indexes are briefly demonstrated in Section 6. Experimental results and conclusions are presented in Sections 7 and 8 respectively.

2. Related work

This dermatological group of disease, i.e. erythemato-squamous, is first studied in [2] and these researchers are also the donors of this dataset. In that study, Guvenir et al. used a classification algorithm called for Voting Feature Intervals (VFI) and they obtained 96.2% accuracy. Guvenir et al. claim another high accuracy of 99.2% in their study with features weighed by a genetic algorithm approach [3]. Another study that makes use of a multi-class SVM strategy [4] is applied to dermatology dataset and a resultant classification accuracy of 98.3% is obtained. The dataset is studied by Nanni [6] for an ensemble classifier algorithm and Nanni obtained 98.3% as his highest accuracy in that work. In another study [7] Polat et al. used fuzzy and k-NN based weighted preprocessing method for a decision tree classifier and they obtained an accuracy of 99.00% as a significant classification performance. In a somewhat different FS strategy, Ozcift et al. used ensemble algorithms to obtain critical features of the dataset and they obtained classification accuracies of 98.64% and 98.91% for SL and BN algorithms [8]. Another study implemented by Karabatak et al. by using an association rule based feature selection technique is presented in [9]. In their study, they used a Neural Network algorithm and they obtained a 98.61% accuracy rate. In a recent study, Xie et al. used SVM combined with a hybrid FS selection strategy and they obtained 98.61% accuracy [10]. The most recent algorithm applied to this dataset is association rule FS based PSO–SVM strategy that provides accuracy of 98.91% [11]. The overall results are summarized in Table 6.

3. Erythemato-squamous diseases data

The differential diagnosis of erythemato-squamous diseases is difficult, since six diseases in this group look very similar with erythema and scaling. With a more careful inspection, it is observed that some patients have the typical clinical features of the disease at the predilection sites (localization of the skin where a disease prefers) while another group has typical localizations.

Patients were first evaluated clinically with 12 features. The degree of erythema and scaling, whether the borders of lesions are definite or not, the presence of itching and Koebner phenomenon, the formation of papules, whether the oral mucosa, elbows, knees and the scalp are involved or not, whether there is a family history or not, are all important features in differential diagnosis.

Some patients can be diagnosed with these clinical features only, however, a biopsy is usually necessary for a correct and definite diagnosis. Skin samples were taken for the evaluation of 22 histopathological features.

Another difficulty for differential diagnosis is that a disease may show the histopathological features of another disease at the beginning stage and may have the characteristic features at the following stages. Furthermore, some samples show the typical histopathological features of the disease while some do not. For a larger discussion of erythemato-squamous diseases, study [3] is recommended.

The *feature names* of the dataset and name of the six diseases are provided in Tables 2 and 3 respectively. Furthermore, we use Table 2 to present a subset of features generated with wrapper algorithms.

```
1: Enter D, F0, M, δ
2: Set Fk with random F0
3: Set K with |Fk|
4: Evaluate J(Fk,D,M) to obtain λ
5: For δ < λ increase K by 1 and go to step 4
6: Else output Fbest = Fk
```

Fig. 1. Filter based FS algorithm.

```
1: Enter D, F0, δ
2: Set Fk with random F0
3: Set K with |Fk|
4: Evaluate J(A(Fk,D)) to obtain λ
5: For δ < λ increase K by 1 and go to step 4
6: Else Fbest = Fk
```

Fig. 2. Wrapper based FS algorithm.

4. Feature selection

One of the core issues in medical data analysis is the so-called *curse of dimensionality*. Redundant features might not only lead to inadequate classification accuracy, but they may also add further difficulty to interpret data [12]. This makes FS an essential step in designing automated diagnosis algorithms. For this aim, reduction of data requires to search an optimal sub set of features that increases classification accuracy. In the literature there are mainly two FS categories: (i) filter methods and (ii) wrapper methods [13]. The algorithmic structure of two FS methods is given in Fig. 1 and Fig. 2. In Figs. 1 and 2, "D is the training dataset with the initial feature set, F_{best} is the optimal feature subset to be selected, and $J(F_k)$ denotes an evaluation function to measure the performance of a feature subset F_k, based on the independent test (M) or the machine learning algorithm (A), respectively" [14].

In filter algorithm the stopping criteria δ might be: (i) subsequent addition or deletion of features does not improve feature subset, (ii) some given bound such as the maximum number of search iterations or the minimum number of features is reached [14].

As Fig. 2 is examined for wrapper approach, it is seen that; for each feature search iteration, the performance of feature subset F_k is evaluated by classification performance of A. This iteration is continued until the predefined criterion δ is met [14].

4.1. Brief discussion of GA wrapper algorithms

The search strategies of the wrapper approaches frequently have heuristic nature rather than being exhaustive. A core classifier for an N feature space is trained 2^N times in an exhaustive search that obviously requires huge amount of computation time [15]. To decrease this computational load, a wrapper must use a search strategy such as greedy search, genetic search or best first search. These search engines, in general, are equipped with a selection fashion such as forward selection or backward elimination [15].

More specifically, GA being a heuristic search algorithm [16], is commonly used to identify relevant features for high dimensional datasets. In genetic algorithms, a potential solution to problem is encoded as a chromosome. This group of chromosomes, i.e. population, is the search space of the algorithm. A fitness function is used to evaluate performance of each chromosome to measure its