Dermatology Diagnosis with Feature Selection Methods and Artificial Neural Network

Shuzlina Abdul-Rahman, Ahmad Khairil Norhan
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
shuzlina@tmsk.uitm.edu.my

Marina Yusoff, Azlinah Mohamed, Sofianita Mutalib
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
{marinay, azlinah, sofi}@tmsk.uitm.edu.my

Abstract—Dermatology or skin disease is one of the popular diseases among other diseases these days. The features similarities between different types of skin diseases make diagnosis of skin diseases very complex. A patient needs dermatologist that has a sound and vast good experience in skin diseases in order to give precise results at the right time. This paper elaborates a prototype with back propagation neural network (BPNN) to assist the dermatologist. This prototype improves expert diagnosis method in term of time efficiency and diagnosis accuracy. The use of two feature selection methods namely Correlation Feature Selection (CFS) and Fast Correlation-based Filter (FCBF) help by providing a smaller number of features with greater accuracy and faster response time. The adjustment of parameter in BPNN gives good performance. The findings show that FCBF method offers the shortest elapsed time and highest result compared to CFS method and the full features with an accuracy of 91.2%.

Keywords: Artificial Neural Network; Dermatology; Feature Selection; Skin disease

I. INTRODUCTION

Skin disease is one of the popular diseases among other diseases these days. Like other diseases, there are varieties of skin disease that is classified as harmful and simple. Simple skin diseases may be easy to treat and cure but for harmful skin diseases, the percentage of uncured patients can possibly be higher. Similarities of symptoms in varieties of skin disease makes diagnosis process for skin disease become very complex. Patient needs dermatologist that has wide and substantial knowledge and experience in skin diseases. These days there are many systems that were developed for medical diagnose purposes. Most of the system focuses on general diseases like heart disease and diabetes but there are less system that focus on skin disease. Example of a general diagnoses system that is well-known and has long been available is MYCIN. The purpose of MYCIN is to diagnose bacterial infections [1]. Skin disease was always taken lightly by human because most of the diseases can be cured in a short period of time. However, once the infected skin is not treated well and ignored, it will soon spread and the person will be infected to a major skin disease that will bring serious effects on skin surface. Presumably it will affect internal organs which may be life-threatening. Varieties of skin diseases with similar features make it hard to get a better result in a short period of time. Thus, a system of diagnosing skin disease would help the expert to get better result in a faster response time.

Guvenir et al. [2] developed an expert system for the differential diagnosis of erythemasquamous disease. Three classification algorithms were used: nearest-neighbour classifier, naive Bayesian classifier and VFI5. The variables are assigned weights to differentiate the irrelevant variables from the relevant variables. Another approach of skin diseases recognition was carried by Antkowiak using skin images [3] whereby the author employed artificial neural network (ANN) and compared with support vector machine (SVM) to recognise skin diseases. ANN has been widely applied in many medical domains [4-6]. In [4], ANN was applied to classify uroflowmeter signals while in [5] ANN was hybridized with fuzzy to perform the automatic system for medical diagnosis. In another study [6] a chest diseases diagnosis was realized by using multilayer, probabilistic, learning vector quantization, and generalized regression neural networks. Other applications of ANN in various domains were reported in [7-9]. In the next subsection, the concept of ANN is briefly explained.

This paper is an extension of previous work that focused on two types of skin diseases [10]. We extend the work of [10] by applying the feature selection methods prior to classification as this would reduce the modelling time and improve the performance. The feature selection methods were applied during the pre-processing stage prior to
classifying the datasets. The remainder of this paper is organised as follows. Section 2 and 3 discuss on the principles of ANN and feature selection methods. In Section 4, the experimental setup is presented and followed by the results discussion in section 5. Finally, in Section 6, the conclusion is given.

II. ARTIFICIAL NEURAL NETWORKS

An artificial neural network (ANN) is a method of computation and information processing that takes advantage of today's technology. Mimicking the processes found in biological neurons, artificial neural networks are used to predict and learn from a given set a data. Neural networks are more robust at data analysis than statistical methods because of their ability to handle small variations of parameters and noise [11-12]. The ANN used in this prototype is a multilayer perceptron (MLP) model with one hidden layer. The prototype was trained using backpropagation algorithm. The algorithm mainly consists of three stages, which are feed forward the input patterns, compute and back propagate the errors and finally updating the weights [12]. The weights are updated after each training set is presented to the networks. The error for a single pattern is calculated as equation 1 while the activation function used in this study is \( y_{\text{sigmoid}} \) and the equation is shown in equation 2.

\[
E_p = \frac{1}{2} \sum_{k=1}^{K} (T_k - O_k)^2
\]  

where \( k = \) output layer size, \( T_k = \) target value, \( O_k = \) actual output

\[
y_{\text{sigmoid}} = \frac{1}{1 + e^{-x}}
\]  

where \( e^{-x} \) is an incoming signal from the network.

III. FEATURE SELECTION METHODS

In real world problems, feature selection (FS) is a must due to the abundance of noisy, irrelevant or misleading features [13]. These phenomena make the FS process critical [13-15] to the data mining success. In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. information theories), FS does not modify the original representation of the features, but merely selects a subset of them or reduces the dimensionality. This allows easy interpretation [16] by a domain expert due to its preservation of the original semantics of the features. Since variables of this dataset are quite large in number, there is an advantage to reduce the number of variables. The effect of reducing variable may give better time efficiency and accuracy. Feature selection methods are proven to be effective in eliminating redundant and irrelevant features, performance of learning algorithm prediction enhanced, better knowledge about data and learning algorithm execution time shorten. Two commonly method for FS are Correlation Features Selection (CFS) [17] and Fast Correlation-Based Filter (FCBF) [15].

1) CFS method ranks high scores to subsets that include features that are highly correlated to class attribute but lowly correlated to each other. This feature selection is available in Weka application in Select Attributes tab, under function Attribute Evaluator “CfsSubsetEval”.

2) FCBF algorithm started with relevance analysis stage, focuses on collection of input aimed at ordering the input variables subject to relevance score, which is the score computer as symmetric uncertainty according to target output. This stage was used to remove less important variables that have ranking score below predefined threshold. Next stage is redundancy analysis. This method chooses major features from the relevant set acquired at relevance analysis. This feature are available in Weka application in Select Attributes tab, first step is to choose “SymmetricalUncertAttributeSetEval” under function Attribute Evaluator then choose “FCBFSearch” under function Search Method.

IV. EXPERIMENTAL APPROACH

The dermatology dataset contains six types of diseases. These skin diseases are psoriasis, seboroeic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis and pityriasis rubra pilaris. These diseases share similar clinical features of erythema and scaling with no much difference. Other than that, a disease may show another disease’s feature at the early stages. The datasets were taken from benchmark dataset website [18] and pre-processed to avoid problems during the modelling creation. The dataset has been divided into two; training set and testing set according to a ratio of 80% for training and another 20% for testing purposes. Therefore, 293 instances out of 366 instances will be placed on training data and the rest which is 73 instances will be placed on testing data. The pre-processing technique includes data normalization, missing value replacement, features interval and features selection method in order to handle the incomplete, noisy and inconsistent data. Since dataset obtained have a few missing values, Weka tool was utilized to fill in the missing value. It is a freeware tool for educational purposes developed at the University of Waikato in New Zealand that composed of a collection of state-of-the-art pre-processing tools and machine learning algorithms [19].
V. RESULTS AND DISCUSSION

This section presents the results and discusses the architecture of ANN and the interface for the prototype.

A. ANN Architecture

Figure 1 shows the ANN architecture that employed the FCBF method. In this architecture, the input size is 16 represents the number of features, which are five clinical features and eleven histopathological features, as shown in Figure 3 and 4. A series of experiments were performed and the final architecture for this study is 16-8-6 (number of input nodes, number of hidden nodes and number of output node, respectively).

B. Prototype Interface

Figure 2 shows the main page of the prototype that has two buttons: the clinical features and the histopathological features. Besides these two buttons, there are another three buttons which are “diagnose” button which is used to diagnose patient symptoms, “exit” button to close the prototype and “reset” button to clear all the previous diagnosis details. The empty space under the result label will display the diagnosis result. Figure 3 illustrates five symptoms of clinical features. Each of these symptom features has four values. The unknown button is checked if the symptoms do not showed any sign while the 1, 2 and 3 indicate the different level of symptoms condition. Level 1 means the condition of symptoms is at severe level, condition level 2 means the condition of symptoms is at the medium level while level 3 indicates the worst condition of the symptoms. The expert has to check all the symptoms. The “submit” button is to store the diagnosis data into file for diagnosis purposes. The “exit” button is to return to the main page. Similar to Figure 3, Figure 4 shows the histopathological features with eleven distinct features.
Figure 4 Diagnosis Section for Histopathological Features

Figure 5 Results Interface

After the diagnosis process finishes, the expert needs to click a “diagnose” button to get the result of the diagnosis as shown in Figure 5. If the expert wants to start a new diagnosis, the “reset” button needs to be clicked in order to clear the previous diagnosis data. The “exit” button is used to close the prototype safely if the expert has finished the diagnosing.

C. ANN Performance

The results are compared in terms of accuracy as well as the number of selected features. Full features means that the entire features was taken as a sample of dataset. FCBF selects the smallest number of features that is 16 input features that consist of 5 clinical features and 11 histopathological features while CFS selected 19 input features that consist of 6 clinical features and 13 histopathological features. CFS selects 19 features out of 34 features that are highly correlated to class attribute while FCBF selects 16 features. Our results showed that both the FCBF and CFS methods select the most highly correlated input features into its classes.

Feature comparisons were performed in order to get the best feature selection method prior to implementing the prototype. A few comparisons were made between the three value of learning rates and epoch for each feature selection method. The value for the learning rates are 0.20, 0.25 and 0.35 and the value for the epochs are 80,000 and 100,000. The best accuracy of feature selection for certain learning rate with certain epoch is further compared in the next stage where different size of hidden layer is exercised. The previous learning rate and epoch are fixed and only nodes in the hidden layer are changed. The nodes of hidden layer were gained by decreasing the original hidden layer nodes by 2 and the other one is by increasing the value of hidden layer nodes by 4.

The last stage of comparison is to compare the average elapsed time for the best result accuracy between CFS, FCBF and Full Features. Table 1 compares the ANN performance with Full Features, CFS and FCBF. As shown in Table 1, both CFS and FCBF gives a superior result compare to Full features. From Table 2, FCBF has recorded the highest accuracy with 91.66% at 80,000 epochs and learning rate of 0.25. CFS has recorded the same accuracy of 91.66% at 100,000 epochs and learning rate of 0.20. Full feature provides the lowest accuracy, 83.33% with 80,000 epochs and the learning rate of 0.20. The highest total average accuracy was given by CFS method with 86.10% at 100,000 epochs. Since the result accuracy between the CFS and FCBF methods are quite similar, the last stage of comparisons was performed. As can be seen in Table 2, FCBF method has recorded the fastest average elapsed time followed by CFS and Full Features that recorded the slowest elapsed time between the three features selections. The final result of the best learning rate is 0.25, the number of epoch is 80,000 and hidden layer is 8. The faster behaviour of the FCBF method is probably due to the design of its algorithm. Unlike the greedy sequential method, the features that have been removed in each round are not considered in the next round and FCBF removes a large number of features instead of only one by one in each round.
TABLE I. ANN PERFORMANCE WITH FULL FEATURES, CFS AND FCBF

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Learning Rate</th>
<th>Result Accuracy (%)</th>
<th>FCBF</th>
<th>CFS</th>
<th>Full Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>80,000</td>
<td>0.20</td>
<td>79.16</td>
<td>83.33</td>
<td>83.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>91.66</td>
<td>83.33</td>
<td>79.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>83.33</td>
<td>87.50</td>
<td>79.16</td>
<td></td>
</tr>
<tr>
<td>Average total accuracy (%)</td>
<td>84.71</td>
<td></td>
<td>84.72</td>
<td>80.55</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.20</td>
<td>75.00</td>
<td>83.33</td>
<td>79.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>87.50</td>
<td>91.66</td>
<td>70.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>83.33</td>
<td>83.33</td>
<td>79.16</td>
<td></td>
</tr>
<tr>
<td>Average total accuracy (%)</td>
<td>81.94</td>
<td></td>
<td>86.10</td>
<td>76.38</td>
<td></td>
</tr>
</tbody>
</table>

#HL: number of hidden layer, FCBF: Fast Correlation Based Filter; CFS: Correlation Feature Selection

TABLE II. ANN PERFORMANCE WITH FULL FEATURES, CFS AND FCBF

<table>
<thead>
<tr>
<th></th>
<th>Learning rate = 0.25</th>
<th>Learning rate = 0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch = 80000</td>
<td>Epoch = 100000</td>
<td>Epoch = 80000</td>
</tr>
<tr>
<td>#HL FCBF #HL CFS</td>
<td>#HL FF</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>79.16 7</td>
<td>87.50 15</td>
</tr>
<tr>
<td>8</td>
<td>91.66 9</td>
<td>91.66 17</td>
</tr>
<tr>
<td>12</td>
<td>87.50 13</td>
<td>79.16 21</td>
</tr>
</tbody>
</table>

#HL: number of hidden layer, FCBF: Fast Correlation Based Filter; CFS: Correlation Feature Selection

TABLE III. THE COMPARISON OF SHORTEST ELAPSED TIME BETWEEN HIGHEST ACCURACY OF CFS, FCBF AND FULL FEATURES

<table>
<thead>
<tr>
<th>Feature size</th>
<th>FCBF</th>
<th>CFS</th>
<th>Full Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>19</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Hidden Layer</td>
<td>8</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>91.66</td>
<td>91.66</td>
<td>87.50</td>
</tr>
<tr>
<td>Epoch</td>
<td>80000</td>
<td>100000</td>
<td>80000</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.25</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td>Average elapsed time (ms)</td>
<td>61</td>
<td>63</td>
<td>73.8</td>
</tr>
</tbody>
</table>

#HL: number of hidden layer, FCBF: Fast Correlation Based Filter; CFS: Correlation Feature Selection

VI. CONCLUSION

The paper discusses the employment of ANN and FS to classify types of dermatology diseases. The use of FS is beneficial in improving the overall performance for diagnosing the skin diseases. The previous work by Yusoff et al. obtained only 87.5% of accuracy while our work achieved 91.2% of accuracy. The proposed solution that incorporated BPNN and FS methods such as CFS and FCBF have shown better performance in terms of accuracy and minimal elapsed times. Future enhancement is to develop a web based system for a better visualization and easy access.

REFERENCES
