A Survey on classification & feature selection technique based ensemble models in health care domain

Garima Sahu
M.Tech (CSE)
Raipur Institute of Technology, (R.I.T.)
Raipur, Chattishgarh, India
garima.sahu03@gmail.com

Rakesh Kumar Khare (H.O.D.)
Department of Information Technology
Raipur Institute of Technology, (R.I.T.)
Raipur, Chattishgarh, India
rakesh_khare2001@yahoo.com

Abstract

There are various ensemble models used in healthcare domain but classification & feature selection technique based models are very effective and useful in diagnosis of skin diseases. The classification technique used in the model determines the scalability, while feature selection technique makes the model quick to diagnose the skin disease. Over the years, various ensemble models have been proposed which are having combination of various classification techniques and some models are based upon feature selection techniques. Some are hybrid models so that accuracy of the model gets increase. Our survey indicates the latest developments in techniques used to build the ensemble model. The combination of techniques varies with the disease so the combination suitable in dermatology can be determined through this survey. The ensemble model can be evaluated on the basis of system parameters such as accuracy, sensitivity, recall etc. The system parameters are represented in terms of percentage, by which ensemble model produces the results with the dataset used.

Keywords- dermatology; expert system; feature selection; classification; sensitivity

1. Introduction

Dermatology is the study related to skin diseases. The process to diagnose the skin diseases is very complex so that sometimes it happens that care providers may not diagnose the disease within time and it may turn out to be skin cancer. The similar features of six categories of skin diseases make it difficult diagnosis. In order to timely identification and accurate results, the concept of ensemble model came into existence so that dependency upon previous experience of care provider can be minimised. To increase the quickness of diagnosis process, verification of

some less important, redundant features available in data collected from patients can be avoided. To fulfil this purpose, a set of minimum number of features is required so that accurate diagnosis results can be produced quickly.

The classification of data can be performed on the basis of features related to diseases, combination of classification techniques are used to build an ensemble model.

The classification techniques works in two ways. First is training mode and other is testing mode. In training mode, the model is used to build using algorithm with the training data related to the domain. In testing mode, the built model is being tested as per the expected results. The deviation of the model from the desired result and the accuracy along with other system features are measured.

This paper gives an overview of the use of data mining techniques such as classification and feature selection technique. Combination of these techniques has been used in various expert systems. These expert systems which are mentioned in this paper are mainly related to health care domain. This paper also conveys the information related to the efficiency of the model proposed by different researchers.

2. Clinical Knowledge Discovery

Clinical knowledge discovery refers to identification of different patterns from raw data related to health care domain and extraction of the relevant information from large databases which consists of patient data. In this process, several steps are present. Those steps involve removal of noise and redundant data. The data which has been
collected from different sources needs to be combined. The data which is specifically related to the analysis task is to be selected or retrieved from large database and the transformation occurs on the application of techniques to utilize the data in specified manner. The whole process is carried down in every expert system with the difference of order and usage of techniques. The care provider may escape themselves from providing the wrong treatment so that expert system may help them in the process of providing effective treatment to the patient.

The associative rules present in expert system closely analyse the features available in the patient data. The effectiveness of the features can be determined through the rules and redundant features can be eliminated to form a set of minimum number of features which are only required for the purpose of diagnosis of the disease. Based upon this methodology, different expert systems are proposed and used by care providers for better treatment. The combination of classification techniques and feature selection methods describe the uniqueness and effectiveness of the proposed expert system.

3. Literature Survey

The clinical decision support systems which are in current use in health care sector and the proposed one are described here for the consolidated view of research and development in the field of clinical decision support system [Table-I].

Marcano-Cedeno et al [1] proposed the expert system for the patients of acquired brain injury. It is based upon the application of Decision Tree (DT), Multi Layer Perceptron (MLP), General Regression Neural Network (GRNN) classification techniques so that efficient and proper care can be provided to patients. Their proposed expert system depicts better results in terms of system features such as accuracy (90.38% with DT, 78.77% with MLP & 75.96% with GRNN), Sensitivity (90.35% with DT, 77.83% with MLP & 75.66% with GRNN) and Specificity (90.62% with DT, 88.26% with MLP & 78.99% with GRNN). The experimental results clearly rank Decision Tree technique as a favourable one to be implemented in the expert system to get better and efficient result.

The results are produced using the patient data stored in the database of hospital named The Institute Guttman.

To get On-demand report of the patients based upon the primary input such as heart rate, number of various blood cells & some other physiological factors collected via different sensors are processed locally and/or at remote location, if required. The infrastructure involves uses wireless network for local analysis and after the determination of index value, if it is greater than a threshold value, the data is to be transferred to the remote system for further accurate analysis to determine the disease and preparation of the reports. In the central database, various techniques such as fuzzy logic and artificial neural network are applied to ensure proper analysis. The two tier architecture is proposed by GennaroTartarisco et al [2]. The report indicates the stress level of the patients and support care providers in decision making by avoiding personal risk.

Use of sentic computing is suggested by Erik Cambria et al [3] to bridge the gap between the text based questionnaire related to symptoms gathered through patients and the use of common sense knowledge. This helps to provide quick medication to the patients.

H. Altay Guvenir et al [4] proposed expert system with voting feature intervals technique which is also known as VFI5. This voting feature interval classification technique uses to identify the new input gathered through the patients by analysing it with the help of available training data. This technique is very effective to increase the efficiency of the diagnosis process because every feature is involved in the voting process separately. This voting is used to determine the predicted class. This accurate and quick classification of features enhanced the productivity of the expert system. This technique involves the projection of features. This projection creates the points of projections, the similarity among these points of projection enables them to be grouped into intervals. An interval is the representation of features with similar class.

This process is followed to select the features from a given set of features collected from...
the patients suffering from disease of dermatology family.

Apart from using a single classification technique in the expert system, Alaa M. Elsayad et al [5] proposed an effective ensemble model which uses three different classification techniques such as C5.0 decision tree technique, multilayer perceptron neural network and linear discriminant analysis. The model is constitute with the optimization effect of individual models involved in the ensemble model so that limitation of individual model can be eliminated so with the help of combination of the techniques, more effective model is built. The decision tree technique divides the pattern in two parts to constitute tree form. The multilayer perceptron neural network presents each input pattern to network and apply the function of weighted sums. Linear discriminant analysis classifies the data patterns belonging to unknown classes based upon available training patterns of already used classes. The proposed ensemble model achieves the accuracy of 98.23% which shows the usability of the build model.

To diagnose thyroid disease (Hyperthyroidism & Hypothyroidism) an expert system is proposed by Ali Keles et al [6]. The expert system uses neuro fuzzy classification method for the purpose of diagnosis of thyroid disease and achieved accuracy of 95.33% by cross validation of ten folds. The system consists of three layer of feed forward neural network. Those layers of neural network represent the input, fuzzy rules and the output. The proposed model has been implemented in an diagnosis application having selection processes to be chosen by the care providers/individuals. It is difficult to diagnose thyroid disease at early stage so that this expert system is very useful so that proper can be arranged.

Ali Keles et al [7] proposed an expert system for another very critical disease which is breast cancer. This expert system is also based upon neuro fuzzy rules. It produces the results related to accuracy, specificity & sensitivity as 96%, 97% & 76%. Kung-Jeng Wang el al [9] try to estimate of survival of breast cancer patients using decision tree classifier along with two new data mining techniques.

A remarkable result has been produced in terms of system accuracy of 99.25% at training stage and 98.99% at testing stage of experiment setup. The accuracy has been achieved by using support vector machine &artificial neural network data mining technique. The proposed expert system by Dinesh K. Sharma et al [8] is very effective to diagnose skin diseases using voting method on the basis of weight assigned to each individual model so that ensemble model may produce better and efficient results. The support vector machine technique creates a hyper plane similar to the surface of decision. In decision surface, the distance between the positive & negative samples gets increased. Through many support vectors, a classification task is classified. This expert system generates good results for other system features i.e.: sensitivity, specificity etc.

Table-I System parameters of proposed classification models for critical diseases

<table>
<thead>
<tr>
<th>Authors/Researchers</th>
<th>Classification technique used</th>
<th>Experiment Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Marcano-Cedeno, Paloma Chausa, Alejandro Garcia, Cesar Caceres, Josep M. Tornos and Enrique J. Gomez [1]</td>
<td>Decision Tree</td>
<td>Accuracy 90.38%, Sensitivity 90.35%, Specificity 90.62%</td>
</tr>
<tr>
<td>Kung-Jeng Wang, BunjiraMakond, Kun-Huang Chen, Kung-Min Wang [9]</td>
<td>synthetic minority oversampling technique, particle swarm optimization &amp; C5.0 decision tree</td>
<td>Accuracy 94.25%, Sensitivity 97.4%, Specificity 94.5%</td>
</tr>
<tr>
<td>H. Altay Guvenir, GulsenDemiroz, NilselIlter [4]</td>
<td>Voting Feature Intervals-5</td>
<td>Accuracy 96.2%</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Methodology</td>
<td>Accuracy</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Alaa M. Elsayad [5]</td>
<td>C5.0 decision tree technique, multilayer perceptron neural network and linear discriminant analysis</td>
<td>98.23%</td>
</tr>
<tr>
<td>A. Marcano-Cedeno, Paloma Chausa, Alejandro Garcia, Cesar Caceres, Josep M. Tormos and Enrique J. Gomez [1]</td>
<td>Multi Layer Perceptron</td>
<td>78.77%</td>
</tr>
<tr>
<td>Ali Keles, Ayturk Keles, Ugur Yavuz [7]</td>
<td>Neuro fuzzy classification method</td>
<td>96%</td>
</tr>
<tr>
<td>Dinesh K. Sharma, H. S. Hota [8]</td>
<td>Support vector machine &amp; artificial neural network</td>
<td>98.99%</td>
</tr>
<tr>
<td>A. Marcano-Cedeno, Paloma Chausa, Alejandro Garcia, Cesar Caceres, Josep M. Tormos and Enrique J. Gomez [1]</td>
<td>General Regression Neural Network (GRNN)</td>
<td>75.96%</td>
</tr>
<tr>
<td>Shelly Gupta, Dharinder Kumar and Anand Sharma [10]</td>
<td>Support vector machine</td>
<td>99.25%</td>
</tr>
<tr>
<td>H.A. Guvenir, N. Emeksiz [12]</td>
<td>Nearest neighbor classifier, naive Bayesian classifier and voting feature intervals-5.</td>
<td>99.2%</td>
</tr>
<tr>
<td>Ridwan Al Iqbal [14]</td>
<td>Attribute Cost Minimized FOCL</td>
<td>90.17%</td>
</tr>
</tbody>
</table>

Shelly Gupta et al [10] has compared various classification techniques using advanced tool such as WEKA [11], Tanagra & Clementine. All the techniques are evaluated on the basis of accuracy percentage and their respective error rate with 10 fold cross validation. The experiment clearly ranks all the available classification techniques which can be implemented to build an expert system for the patients of diabetes.

H.A. Guvenir et al [12] not only propose an expert system but implement it into a visual tool which can be used for the purpose of diagnosis of complex skin diseases. It uses three classification techniques named nearest neighbor classification, naive bayesian classifier with normal distribution method and the advanced classification technique, voting feature intervals-5. The nearest neighbor classification determines the closeness of unknown sample to a group of similar patterns. The closeness is determined by calculating the Euclidean distance between them (Jiawei Han, Micheline Kamber [13]). The conditional probability of each feature is calculated in naive bayesian classification technique and in VFI-5, score of each feature is being calculated by giving assigned weightage to them, on the basis of voting, class is determined for each feature.

Ridwan Al Iqbal [14] suggested a hybrid approach for the formation of clinical decision support system having the qualities of knowledge based expert system as well as data mining based expert system. It reduces the dependency upon the reliability of the training data so that more accuracy can be achieved. In this hybrid approach, inference engine will map new sample with the existing rules derived from the knowledge of domain experts and if the correct matching of attribute is not available, expert system will try to classify with the available training data and model will accurately produce the expected results.

The feature selection methodology proposed by Ahmed K. Farahat et al [15] is unique because it uses the recursive method of unsupervised learning environment and implements it into the supervised learning environment. It selects the features based upon criteria of measurement of error in reconstruction of matrix of data collected from the subset of the features which
are already selected. They propose the recursive formula for the criteria along with the effective greedy algorithm.

Diagnosis time has been reduced in a great amount with the application of feature selection technique before the application of classification technique or building the model because there can be many irrelevant and redundant features which are not required to be incorporated within the formation of the model for the detection of skin diseases. The feature selection techniques used by Shuzlina Abdul-Rahman et al [16] are correlation feature selection and fast correlation based filter. Correlation feature selection ranks all the features on the basis of their respective co relevancy with the class attribute and fast correlation based filter method directly remove the features which score less than a threshold value and provide the ranks to them. After feature selection process gets completed, the classification technique of artificial neural network is applied. The neural network used in this model is of multilayer perceptron having one hidden layer. Through back propagation neural network, better performance can be achieved by adjusting the parameters. This model achieved the accuracy of 91.2% which is a significant number for the expert system.

4. Conclusion from Literature Survey
This survey paper elaborates the key combination of classification techniques and the effect of using more than one technique. This paper also aims the development in the feature selection methods which definitely accelerates the efficiency of the expert systems of health care domain. Many complex diseases can be detected automatically at early stage with the help of these artificial intelligence systems. This paper also gives an idea to the researchers about the effect of data mining techniques over the dataset related to the patients of different diseases so that classification and feature selection technique can be chosen easily for specific disease. The research dataset available at UCI repository [17] is very reliable because most of the expert systems have been proposed with the experimental result upon it. This survey paper embarks the development and automation in health care domain. It shows good signs for further research and development in the health care sector.

5. REFERENCES


