Sci.Int.(Lahore),26(5),1667-1671,2014 ISSN 1013-5316; CODEN: SINTE 8

A DIAGNOSTIC MODEL BASED ON HYBRID FEATURES SELECTION METHOD FOR THE DIAGNOSIS OF CLINICAL DISEASES

Fatihah Mohd[1], Noor Maizura Mohamad Noor[1], Zainab Abu Bakar[2], Zainul Ahmad Rajion[3], Hasni Hassan[4]
[1]School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia.
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknology MARA (UiTM),
40450 Shah Alam, Selangor, Malaysia.
[3]School of Dental Sciences, Universiti Sains Malaysia (USM),
16150 Kubang Kerian, Kelantan, Malaysia.
[4]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA),
22200 Besut, Terengganu, Malaysia.
Email: mpfatihah@yahoo.com, maizura@umt.edu.my

ABSTRACT : *This paper developed a diagnosis model based on hybrid feature selection method to diagnose erythemato-squamous diseases. Our hybrid feature selection method, named HCELFS (Hybrid Correlation Evaluator and Linear Forward Selection), combines the advantages of filters and wrappers to select the optimal feature subsets from the original feature set. In HCELFS, Correlation Attribute Evaluator acts as filters to remove redundant features and Linear Forward Selection with some machine learning algorithms acts as the wrappers to select the ideal feature subset from the remaining features. Several experiments using WEKA had been conducted, utilizing 10 fold cross validations. The experimental results with erythemato-squamous diseases data set demonstrate that our proposed model has a better performance than some well-known feature selection algorithms with optimal classification accuracy with no more than 16 features for erythemato-squamous diseases.*

KEYWORDS: Correlation attribute evaluator , Diagnosis, Erythemato-squamous diseases, Features selection, Linear forward selection, SMOTE

1.0 INTRODUCTION

Differential diagnosis of diseases is a complicated dilemma in many clinical fields. There are two main problems in diagnosis of diseases. First, most of clinical diseases share the same clinical features and scaling. Generally, a biopsy is taken for the diagnosis of diseases. However the diseases often share many histopathological features as well. The second problem is, one disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages [1, 2]. One of the diseases sharing this dilemma is erythemato-squamous. There are six groups of erythemato-squamous diseases (ESD) including psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis and pityriasis rubra pilaris [3, 4].

The difficulty to diagnose ESD has attracted some experts to study the problem from the perspectives of both medical and computer science. A variety of machine learning algorithms in artificial intelligence such as features selection (FS) methods and classifications are usually applied in the diagnosis of clinical diseases. Both feature selection processes and classification techniques are capable of producing the most relevant features to build an efficient classifier. In addition, they can also remove noisy and redundant features to obtain a classification with higher accuracy. Examples of common classification methods include Bayesian Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Networks (ANN). These efficient methods are able to aid doctors in making informed decision of diseases based on the features obtained from the classification. This paper reviews some related studies in the attempt to discover an efficient feature selection method to build the diagnostic model for ESD. We also suggested our new hybrid feature selection methods to diagnose the diseases using popular classification techniques such as Naïve Bayes (NB), Multilayer Perceptron (MLP), SVM and KNN.

Clinical data sets are usually not balanced. Data sets are unbalanced when at least one class is represented by only a small number of training examples (called the minority class) while other classes make up the majority. In this scenario, classifiers can have good accuracy on the majority class but very poor accuracy on the minority class(es) due to the influence that the larger majority class has on traditional training criteria. Most classification algorithms aim to minimize the error rate and the percentage of incorrect prediction of class labels where most of the times, the difference between types of misclassification errors are ignored. In particular, they implicitly assume that all misclassification errors have equal cost [5]. To overcome this problem, we propose preprocessing on the imbalanced data set before the features selection stage. In this study, we integrate the Synthetic Minority Oversampling technique (SMOTE) algorithm in our diagnostic model to resolve the problem of imbalance data set.

The remaining of the paper is organized as follows: The following section provides a brief description about the erythemato-squamous diseases, features selection methods used in this study, the proposed diagnostic model for ESD, and our hybrid FS algorithms. Next, the experimental findings are elaborated in results and discussion section and finally the concluding remarks in the last section address further research issues in this area.

2.0 MATERIAL AND METHODS

This section first describes the erythemato-squamous diseases data set, and then elaborates the features selection methods used in this study. The development of the diagnostic model is also explained in this section together

with the proposed hybrid features selection method for ESD diagnosis.

2.1. Erythemato-Squamous Diseases (ESD) Data Set

The experimental work in this study used ESD data set from UCI Machine learning depository (http://archive.ics.uci.edu/ml). This database has 34 attributes, 33 of which are linear valued and one of them is nominal. Patients were first evaluated clinically with 12 features. Then skin samples were evaluated for 22 histopathological features under a microscope analysis. The numbers of instances were 366, while the numbers of attributes were 34.

2.2. Features Selection Methods (FS)

In this study, feature selection for high-dimensional data were performed using WEKA with 10 fold cross validations. The main objective off feature selection functions is to find the most significant features by removing features with little or no predictive information. The Following are the functions used for features selection in this study:

i. Correlation Attribute Evaluator (CAE). This algorithm evaluates the worth of features by measuring the correlation between the corresponding features and the class. Nominal features are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal feature is derived via a weighted average.

ii. CFS Subset Evaluator. This algorithm evaluates the worth of a subset of features by considering individual predictive ability of each feature along with the degree of redundancy between them. The Correlation Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function.

ii. Ranker. The purpose of a Ranker is to rank features by their individual evaluations. It is used in conjunction with feature evaluators.

iv. Linear Forward Selection (forward) (LFS). Linear Forward Selection function is extension of Best First method. This technique reduces the number of features expansions in each forward selection step. The search direction of this technique can be forward or floating forward selection (with optional backward search steps).

2.3. Diagnostic Model for ESD

Figure 1 shows the process flow of diagnostic model for ESD. It begins with a number of data set with a set of features supplied as input to the system. ESD data is imbalance in nature hence the data needs to be preprocessed prior to the next stage of processes. Data set is unbalanced when at least one class have only a small number of instances (called the minority class) while other classes is a majority (with a large number of instances). In this scenario, classifiers of the majority class usually have good accuracy while the minority class(es) has/have very poor accuracy. In this study, Synthetic Minority Oversampling Technique (SMOTE) algorithm was applied to resolve the problem of imbalance data set during the preprocessing stage. The next stage is features selection. Three features of selection methods are applied to original features; it began with no features selection, then the correlation with ranker and lastly the features subset from the correlation ranker is applied again to linear forward selection method. All the features subset selected from FS methods are trained with classification algorithm. The performance accuracy of classification algorithms will verify the best selected features subset for clinical diagnosis.

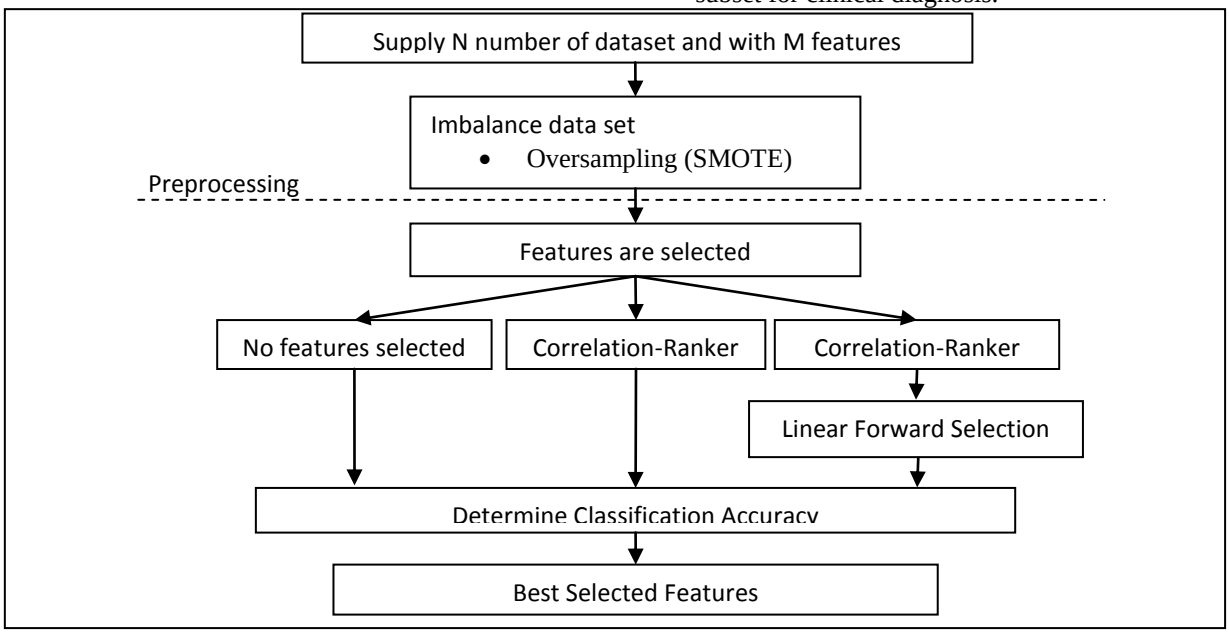

Figure 1: Diagnostic model for ESD

2.4. Hybrid Correlation Evaluator and Linear Forward Selection (HCELFS)

There are six stages in our proposed hybrid feature selection method. The method is called HCELFS, where it combines hybrid CAE with ranker and CFS Subset Evaluator with LFS to select the optimal features subset from the original feature set. In this diagnostic model, we applied the correlation attribute evaluator method to filter the relevant features which resulted in reduced features subset. From this subset CFSs subset evaluator with LFS searched for the most

Sci.Int.(Lahore),26(5),1667-1671,2014 ISSN 1013-5316; CODEN: SINTE 8

relevant features resulting in optimum feature set used for diagnosing ESD. Our HCELFS algorithm selected the optimum features as follows:

Input: Data set with all features (35 features)
Output: Data set with optimum features
Step 1: Rank the relevant features from highest to lowest using correlation attributes evaluator method
Step 2: Remove 1/3rd of the features from the feature set which has lowest ranking rate
Step 3: Add the remaining 2/3rd feature set in the queue to the next steps
Step 4: Reduce the features subset to remove redundant and irrelevant features
Step 5: Select the most relevant features using Linear Forward Selection
Step 6: The features in the queue is the required optimum feature subset necessary for the diagnostic model

3.0 RESULTS AND DISCUSSION

In this study, the ESD data were categorized into six classes. There were 112 instances in the majority class (psoriasis), 72 for lichen planus, 61 for seboreic dermatitis and the other three classes (pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris) falls under the category of minority class with the number of instances less than 60. For the training set, 10-fold cross-validations were used. The minority class was over-sampled at 100%, 200%, 300%, and 400% of its original size. Table 1 shows the result of re-sampling an imbalance ESD using SMOTE. After the process of over sampling, the number of instances became 660 instead of the original 366 instances. Over sampling has resulted in an almost balanced class distribution among the minority and the majority classes as follows: Minority class-pityriasis rosea (14.85%), chronic dermatitis (15.76%) and pityriasis rubra pilaris (12.12%) and Majority class-psoriasis (16.97%), seboreic dermatitis (18.48) and lichen planus (21.82%).

The experiments of features selection against ESD data set were conducted using WEKA with 10-fold cross validations. The algorithm started with 34 features and 660 instances. Using Correlation Ranking Filter, the algorithm ranked 34 features namely 33, 27, 29, 21, 6, 12, 25, 8, 9, 22, 16, 20, 14, 15, 28, 4, 10, 5, 3, 31, 7, 30, 24, 11, 26, 23, 2, 19, 18, 34, 13, 17, 1, and 32. We removed 11 features namely 11, 26, 23, 2, 19, 18, 34, 13, 17, 1, and 32 with lowest ranking which then resulted with 23 features. Next, the subset method removed the redundant and irrelevant 7 features namely, 27, 6, 12, 8, 10, 30, and 24. Finally, the algorithm ended with 16 features namely 33, 29, 21, 25, 9, 22, 16, 20, 14, 15, 5, 28, 4, 3, 31 and 7 as the optimum features set. The implementation of hybrid the features selection methods in this study started with no features selection (FS0), CAE with Ranker (FS1), and then combined FS1 with CfsSubset Evaluator with LFS (FS2).

Four different machine learning algorithms were applied to classify the ESD data set with three features selection methods and optimum features selected by the proposed hybrid algorithm, HCELFS. The four algorithms used were NB, MLP, KNN and SVM. In order to evaluate the efficiency of the FS methods, performance measure of accuracy were considered. The measures were compiled by the classification accuracy (%) = (TP+TN) / (TP + FP + FN +TN). Table 2 shows the results for the classifier by three features selection methods with and without oversampling method, SMOTE. The empirical comparison shows that without using SMOTE, the accuracy of the entire classifier algorithm used for the ESD data set didn't improve. However, using oversampling (SMOTE), the results for three features selection methods with four classifiers show

Table 1: Balanced class distribution for ESD by applying SMOTE

Class	Class name	# of instances	%	# of instances with SMOTE	%
1	Psoriasis	112	30.60	112	16.97
2	Seboreic dermatitis	61	16.67	122	18.48
3	Lichen planus	72	19.67	144	21.82
4	Pityriasis rosea	49	13.39	98	14.85
5	Chronic dermatitis	52	14.21	104	15.76
6	Pityriasis rubra pilaris	20	5.46	80	12.12
	Total	366		660	

Table 2: Performance accuracy for classification algorithms on ESD data set without and with SMOTE

	without SMOTE			with SMOTE		
Algorithm	**FS0**	**FS1**	**FS2**	**FS0**	**FS1**	**FS2**
NB	97.541 2.459	96.1749 3.8251	96.4481 3.5519	98.4848 1.5752	98.1818 1.8182	98.0303 1.9697
MLP	97.541 2.459	94.8087 5.1913	95.3552 4.6448	98.1818 1.8182	97.5758 2.4242	98.0303 1.9697
SVM	97.2678 2.7322	95.6284 4.3716	96.4481 3.5519	98.6364 1.3636	98.6364 1.3636	98.6364 1.3636
KNN	95.3552 4.6448	93.9891 6.0109	94.8087 5.1913	99.0909 0.9091	98.1818 1.8182	98.1818 1.8182

Table 3: Comparison on classification accuracies from the literature using the dermatology data set

Author, year	**Method**	**Accuracy %**
Übeylı and Güler, 2005 [6]	ANFIS	95.50
Nanni, 2006 [7]	LSVM based on random subspace (RS)	97.22
Übeyli, 2008 [8]	Multiclass SVM with the ECOC	98.32
Übeyli, 2009 [9]	Combined neural network (CNN)	96.71
Lekkas and Mikhailov, 2010 [10]	Evolving fuzzy classification	97.55
Xie et al., 2010 [11]	IFSFFS	97.58
Xie and Wang, 2011 [12]	IFSFS with SVM	98.61
Aruna et al., 2012 [13]	A hybrid FS based on IGSBFS with SVM	98.36
Menai and Altayash, 2014 [14]	Boosting decision trees	96.72
This study	Hybrid CAE and LFS (HCELFS) with SVM	98.64

T

that the features selected by the hybrid algorithm contributed to improved accuracy of the entire classifier algorithm used for the ESD data set. The accuracy improves from 96.45% to 98.03% for NB, 95.36% to 98.03% for MLP, 96.45% to 98.64% for SVM and 98.81% to 98.18% for KNN. Findings from Table 2 also shows that the new HCELFS (F2) has the highest classification accuracy performance using SVM algorithm, with an average accuracy of 98.64%.

Table 3 summarizes classification accuracies of all available methods from literature for ESD data set. Our method using SVM with HCELFS outperforms the other available methods with the optimal classification accuracy of 98.64% to diagnose ESD with the size of the selected feature subset of 16 features. The reasons for this good performance are due to two aspects. One is the oversampling method, SMOTE (used to balance between minority class and majority class data); the other is the hybrid features selection algorithms that have been applied to obtain the best result in the selected features subset.

4.0 CONCLUSION

Features selection is a very common and useful process in reducing dimensionality to reduce unrelated data and increase the results to, we investigated features of selection methods to be applied in the diagnosis of ESD using machine learning classification algorithms. HCELFS, a diagnostic model with SMOTE at a preprocessing stage; integrated with a hybrid FS method to diagnose the group of ESD has showed an increase in classification accuracy. Being a hybrid feature selection method, HCELFS combines CAE which acts as a filter and SBFS which acts as the wrapper to select the ideal feature subset from the remaining features. The experimental results with ESD data set demonstrated that HCELFS method has a better performance than some well-known feature selection algorithms with an accuracy of 98.64% where it obtained optimal classification accuracy with 16 features from a set of 34 features. The optimal feature subset were obtained and trained with various data mining algorithms such as NB, MLP, KNN, and SVM to diagnose the stage of ESD. Having obtained a promising result in this direction, the focus on future studies is to consider proposing a hybrid algorithm with various datasets using other data mining classifiers.

ACKNOWLEDGEMENTS

This study has been supported in part of the Exploratory Research Grant Scheme (ERGS) 600_RMI/ERGS 5/3 (3/2011) under the Malaysia Ministry of Higher Education (MOHE) and Universiti Teknologi MARA (UiTM) Malaysia. The authors would like to acknowledge all contributors, technical members at Hospital Universiti Sains Malaysia (HUSM) who have provided their assistance in the completion of the study and anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

REFERENCES

[1] J. He, X. Liu, E. Krupinski, and G. Xu, Ed., Novel Hybrid Feature Selection Algorithms for Diagnosing Erythemato-Squamous Diseases, ser. Health Information Science, Springer Berlin Heidelberg, 2012, vol. 7231.

[2] B. Karlk and G. Harman, "Computer-aided software for early diagnosis of eerythemato-squamous diseases," in *Proc. Electronics and Nanotechnology (ELNANO), 2013 IEEE XXXIII International Scientific Conference*, 2013, p. 276-279.

[3] H. A. Güvenir, G. Demiröz, and N. İlter, "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals," *Artificial Intelligence in Medicine*, vol. 13, pp. 147-165, 1998.

[4] H. A. Güvenir and N. Emeksiz, "An expert system for the differential diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 18, pp. 43-49, January. 2000.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, pp. 321-357, 2002.

[6] E. D. Übeylı and İ. Güler, "Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems," *Computers in Biology and Medicine*, vol. 35, pp. 421-433, 2005.

[7] L. Nanni, "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases," *Neurocomputing*, vol. 69, pp. 842-845, 2006.

[8] E. D. Übeyli, "Multiclass support vector machines for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 35, pp. 1733-1740, 2008.

[9] E. D. Übeyli, "Combined neural networks for diagnosis of erythemato-squamous diseases.," *Expert Systems with Applications,* vol. 36, pp. 5107-5112, 2009.

[10] S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases," *Artificial Intelligence in Medicine,* vol. 50, pp. 117-126, 2010.

[11] J. Xie, W. Xie, C. Wang, and X. Gao, "A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythemato-squamous diseases," *Journal of Machine Learning Research-Workshop and Conference Proceedings 11,* vol. pp. 142-151, 2010.

[12] J. Xie and C. Wang, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases," *Expert Syst. Appl.,* vol. 38, pp. 5809-5815, 2011.

[13] S. Aruna, L. V. Nandakishore, and S. P. Rajagopalan., "A hybrid feature selection method based on IGSBFS and naive bayes for the diagnosis of erythemato-squamous diseases," *International Journal of Computer Applications* vol. 41, pp. 13-18, March. 2012.

[14] M. Ali, J.-S. Pan, S.-M. Chen, and M.-F. Horng, Ed., Differential Diagnosis of Erythemato-Squamous Diseases Using Ensemble of Decision Trees, ser. Modern Advances in Applied Intelligence, Springer International Publishing, 2014, vol. 8482.