Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree

Keivan Maghooli1, Mostafa Langarizadeh2, Leila Shahmoradi3, Mahdi Habibi-koolae3,5, Mohamad Jebraeily6, and Hamid Bouraghi7

ABSTRACT

Introduction: Differential diagnosis of Erythmato-Squamous Diseases (ESD) is a major challenge in the field of dermatology. The ESD diseases are placed into six different classes. Data mining is the process for detection of hidden patterns. In the case of ESD, data mining help us to predict the diseases. Different algorithms were developed for this purpose. Objective: we aimed to use the Classification and Regression Tree (CART) to predict differential diagnosis of ESD. Methods: we used the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. For this purpose, the dermatology data set from machine learning repository, UCI was obtained. The Clementine 12.0 software from IBM Company was used for modelling. In order to evaluation of the model we calculate the accuracy, sensitivity and specificity of the model. Results: The proposed model had an accuracy of 94.84% (Standard Deviation: 24.42) in order to correct prediction of the ESD disease. Conclusions: Results indicated that using of this classifier could be useful. But, it would be strongly recommended that the combination of machine learning methods could be more useful in terms of prediction of ESD.

Key words: classification; Classification and Regression Tree; classifier; data mining; dermatology; Erythmato-Squamous Diseases.

1. INTRODUCTION

In the field of dermatology, differential diagnosis of Erythmato-Squamous Diseases (ESD) is difficult. The six important types of ESD are including psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis and pityriasis rubra pilaris. All of them have similar clinical and histopathology features such as erythema and scaling which makes it difficult to detect them even with biopsy (1, 2). However, biopsy is necessary to achieve a correct diagnosis (3). Usually, clinicians detects these diseases using some clinical features, such as the degree of erythema and scaling, border of lesion, the presence of itching, koebner phenomenon, the formation of papule, involvement of oral mucosa, elbows, knees and the scalp and family history (1).

Data mining defines as discovery of hidden patterns in large volumes of data (4) and widely used in the health care. Using data mining techniques can be answered clinical questions in terms of diagnosis and treatment of diseases (5). The major data mining techniques which used in the health care are including discovering techniques such as clustering and association rules, and prediction techniques such as classification and regressions (5). Data mining algorithms used in the health care are commonly determined based on the information needs (6). This study aimed to investigate the CART algorithm to classify the ESD.

Related works

In this area, several studies have been done using different data mining methods. In all studies, the same dermatology data set is used.

The first attempt was done by Güvenir, Demiröz (1) and new classifier algorithm, the voting feature intervals-5 algorithm was developed. For performance evaluation, the accuracy was measured. The achieved accuracy of the classifier was 96.2%. Menai and Altayash (7) investigated the performance of boosting decision tree as an ensemble strategy for the diagnosis of ESD. The result illustrated that the ensemble of unpruned decision tree had a better accuracy (96.72%) than other
methods such as genetic algorithm (GA) and k-means clustering in other studies.

Xie, Xie (8) Developed a Support Vector Machine (SVM) model with a novel hybrid feature selection method, called Improved F-score and Sequential Forward Floating Search (IFSFFS) which was a combination of Sequential Forward Floating Search (SFFS) and Improved F-score (IF) to carry out the optimal feature subset selection. The IF and SFFS based on SVM are evaluation criteria for filters and wrappers, respectively. They calculated the accuracy of training and testing data set for different splits. The average training and testing accuracy was 99.29% and 97.58%, respectively. Olutunji and Arif (9) proposed and implemented a novel identification model for diagnosis of ESD based on Extreme Learning Machine (ELM). The new model compared with classic Artificial Neural Networks (ANN). ELM had some advantages, such as higher learning speed, the best performance and ease of implementation. Results showed that the ELM had greater degree of performance than classic ANN in both training and testing data set.

Ravichandran, Narayananurthy (10), used a Fuzzy Extreme Learning Machine (FELM) method to purpose ESD classification. FELM consists of two methods: Fuzzy Logic and ELM. The achieved accuracy of the FELM was 92.84%, which was better than other methods in terms of accuracy.

Accuracy of C4.5 classifier calculated as 84.48% in the study of Polat and Güneş (11). They combined C4.5 method with one-against-all method and accuracy of 96.71% was achieved. In (12) an expert system based on three classifier method including decision tree, fuzzy weighted pre-processing and k-NN based weighted pre-processing has been designed in order to help physicians for diagnosis of ESD. While, an ensemble of SVM based on random subspace (RS) and feature selection were employed for diagnosis of ESD by Nanni (3). It is stated that classifiers using different features offers complementary information about the classifiable patterns. Abdi and Giveki (13) Developed a diagnosis model for automatic detection of ESD, using the particle swarm optimization (PSO) method and SVM method based on Association Rules (AR). The accuracy of 98.91% was achieved for AR-PSO-SVM model. Übeyli (14) used the SVM to classify the ESD.

Recurent neural network (RNN) and multilayer perceptron neural network (MLPNN) were compared with SVM. The accuracy of SVM, RNN and MLPNN was 98.32%, 96.65% and 85.47% respectively. In another study Übeyli and Guler (15) the Adaptive neuro-fuzzy inference system (ANFIS) was used as a classifier and the achieved accuracy reported as 85.5%.

Decision tree is a classification and prediction technique of data mining, which is faster than Neural Networks (NN), and because of good understanding is used in the field of health care and medicine (16). Decision tree learning is widely used and is a practical method for inductive inference (12). Decision tree algorithm applied in supervised learning. Best known decision tree algorithms are the following: Classification and Regression Tree (CART), C4.5, Iterative Dichotomizer 3 (ID3), supervised learning in Quest (SLIQ) and Scalable classifier for Data Mining (SPRINT)(17).

2. METHODOLOGY

In this study, the Cross Industry Standard Process for Data Mining (CRISP-DM) is used. The CRISP-DM methodology is described in terms of a hierarchical process model. The CRISP-DM reference model defines data mining in six phases (18). Each phase of the study described as following:

Business understanding

As mentioned in the introduction section, the six phenotypes of Erythmato-squamous Diseases are the major challenge in dermatology. For instance, the psoriasis is the most common disease of ESD which approximately, 2 to 3% of population are affected. Due to overlapping between sign and symptom of ESD, differential diagnosis is very difficult (7). These diseases are frequently seen in the outpatient department of dermatology. All of ESD disease are same in terms of erythema and scaling, but when examine carefully some patient have the typical clinical feature of disease (l). Automatic detection of these group of disease could help physician for decision making (7).

Data understanding

In this study, a standardized dermatology data set from machine learning repository or UCI, developed by University of California, School of Information and Computer Science was used(19). This data set was prepared by N. Ilter from Gazi University and H.A. Guvenir from Bilkent University(1, 7).

The dermatology data set contains 366 samples and 34 at-

| classes | Features |
|---------|----------|
| psoriasis | erythema, melanin incontinence |
| seborrheic dermatitis | scaling, eosinophilia in the infiltrate |
| lichen planus | definite borders, PNL infiltrate |
| pityriasis rosea | itching, fibrosis of the papillary dermis |
| chronic dermatitis | koebner phenomenon, exocytosis |
| pityriasis rubra pilaris | polygonal papules, acanthosis |
| follicular papules | hyperkeratosis |
| oral mucosal involvement | parakeratosis |
| knee and elbow involvement | clubbing of the rete ridges |
| scalp involvement | elongation of the rete ridges |
| family history, (0 or 1) | thinning of the supra-papillary epidermis |
| Age (linear) | spongiform pustule |
| muro microabscess |
| focal hypergranulosis |
| disappearance of the granular layer |
| vacuolization and damage of basal layer |
| spongiosis |
| saw-tooth appearance of rete |
| follicular horn plug |
| perifollicular parakeratosis |
| inflammatory mononuclear infiltrate |
| band-like infiltrate |

Table 1. Clinical and histopathological features of dermatology data set and the classes
tributes. As Table 1 showed, of 34 attributes, 12 are clinical features and 22 are histopathological features. From the data set features, age and family history are continuous and ranged between 0-1, respectively. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3 which, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values (1, 7).

**Data preparation**

In this phase, the quality of data was surveyed. There are 2.2% of missing data in the data set which it was for age attribute. Missed data replaced with real values, using mean frequency. To ensure that there is not outlier data, the histogram chart was used. For the purpose of decision tree classifier, it is needed to discretize the age attribute. The age attribute discretized using equal frequency into four split, using Information Gain measure.

**Modelling**

Several methods presented to classify Erythmato-Squamous diseases. Such methods including, k-means, genetic algorithm, support vector machine, artificial neural network, clustering, and decision tree. Since, decision trees are the most common used and practical method for induction inference, it is used for classification of ESD.

Decision tree consist of a collection of decision nodes, which connected with branches. A decision tree begins with a root node and ends with leaf nodes (20). Learning tree can represented as sets of if-then rules to improve human readability. The aim of decision tree learning is recursively partition data into subgroups (7, 12). In decision tree, it is needed to discretization of attributes and must be consider during data preparation phase (20).

As mentioned above, there are several algorithm to construct decision tree. In this study, the Classification and Regression Tree (CART) algorithm was used.

The CART algorithm, introduced by Breiman et al in 1984, produced a binary decision tree. Larose (20) Described CAR.D algorithm as the following:

CART recursively partitions the records in the training data set into subset of records with the similar values for the target attributes. The tree grows and search all current attributes to select the optimal split according to (1), which is a measure of the “goodness” of a candidate split $s$ at node $t$ where $t_l$ and $t_r$ are left and right child node of $t$, respectively. $P_l$ and $P_r$ are the number of records at $t_l$ and $t_r$ per total records in the training data set, respectively. Also, the optimal split is whichever split maximizes this measure over all possible splits at node $t$ (20).

For the purpose of modeling, the Clementine 12.0 software from IBM Company was used in this study.

**Evaluation**

The dermatology data set partitioned into three parts. One part for training of classifier that consisted 60% of the whole data set (213 cases). And then, 30% for the purpose of testing (111 cases) and 10% for evaluation of the model (42 cases).

Total accuracy of classifier, sensitivity and specificity measures was used for performance evaluation, where defined as following (14, 15).

**Deployment**

In this phase, an appropriate report from the model is provided, according to the objectives of the study.

**3. EXPERIMENTAL RESULTS**

In this study, the data set consisted of 366 samples, which 213 samples used for training of the model. The important attributes in differential diagnosis of ESD are: clubbing of the rete ridges, polygonal papules, fibrosis of the papillary dermis, koebner phenomenon and follicular papules, respectively. After training the data set, the test data set was used for diagnosis of ESD. Table 2 below indicates confusion matrix for testing data set, which show the number of cases that classified as seborrheic dermatitis by the model. For lichen planus, the model classified one case as seborrheic dermatitis and another case as pityriasis. Also, four cases of pityri-
Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree

After that, accuracy, sensitivity and specificity calculated for performance evaluation of the model. The “seborrheic dermatitis” and “pityriasis rubra pilaris” had 100% sensitivity. Also, the “lichen planus”, “chronic dermatitis” and “pityriasis rubra pilaris” have 100% specificity. Total accuracy of the model was 93.69% (Standard Deviation 24.42). Hence, the CART algorithm can predict differential diagnosis of ESD with 93.69% accuracy.

4. DISCUSSION

As mentioned in the literature review, many studies conducted to determine differential diagnosis of ESD, using different methods. The present study was designed to determine differential diagnosis of ESD, using CART classifier.

In the current study, the accuracy of classifier were obtained 95.31% and 93.69% for training and testing data set, respectively. These finding of the study indicate that the using of CART classifier do not present an appropriate accuracy in comparison with other classifiers.

These results differ from some published studies. Table 3 illustrate the comparison of this study with those study that used different classification methods for differential diagnosis of ESD using dermatology data set.

According to Table 3, studies that used the ensemble methods to classify ESD, achieved a better accuracy than other studies that used a simple method. For instance, in the study of Abdi and Giveki (13) and in the study of Xie, Xie (8), ensemble of SVM and other methods gained a better accuracy than SVM alone.

5. CONCLUSION

This study set out with the aim of prediction of differential diagnosis of ESD using classification and regression tree (CART). Although, the CART classifier achieved a better accuracy than some methods such as ANN and C4.5, but it is strongly recommended that ensemble methods should be used to classify differential diagnosis of ESD. Further experimental investigations are needed to conduct a prediction model of ESD, using CART classifier in combination with other methods.

• Conflict of interest: none declared.

REFERENCES

1 Güvenir HA, Demiröz G, Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artificial Intelligence in Medicine. 1998; 13(3): 147-65.
2 Aruna S, Nandakishore L, Rajagopalan S. A Hybrid Feature Selection Method based on IGSSBFS and Naïve Bayes for the Diagnosis of Erythema-Squamous Diseases. International Journal of Computer Applications (0975–8887). 2012; 41(7).
3 Nanni L. An ensemble of classifiers for the diagnosis of erythematous-squamous diseases. Neurocomputing. 2006; 69(7): 842-5.
4 Han J, Kamber M. Data mining : concepts and techniques. 3rd ed. Burlington, MA: Elsevier, 2011.
5 Raghuapathi W. Data mining in health care. Healthcare Informatics: Improving Efficiency and Productivity. 2010: 211-23.
6 Hemalatha M, Megala S. Mining Techniques in Health Care: A Survey of Immunization. Journal of Theoretical and Applied Information Technology (JATIT). 2011; 25(2).
7 Menai MEB, Altayash N. Differential Diagnosis of Erythemato-Squamous Diseases Using Ensemble of Decision Trees. Modern Advances in Applied Intelligence: Springer; 2014: 369-77.
8 Xie J, Xie W, Wang C, Gao X, editors. A Novel Hybrid Feature Selection Method Based on IFSFFS and SVM for the Diagnosis of Erythemato-Squamous Diseases. WAPA, 2010.
9 Olatunji SO, Arif H. Identification of Erythemato-Squamous skin diseases using extreme learning machine and artificial neural network. ICTACT Journal of Softw Computing. 2013; 4(1): 627-32.
10 Ravichandran K, Narayananmurthy B, Ganapathy G, Ravalli S, Sindhura J. An efficient approach to an automatic detection of erythematous-squamous diseases. Neural Computing and Applications. 2014; 25(1): 105-14.
11 Polat K, Güneş S. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications. 2009; 36(2):1587-92.
12 Polat K, Güneş S. The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythematous-squamous diseases. Digital Signal Processing. 2006; 16(6): 922-30.
13 Abdi MJ, Giveki D. Automatic detection of erythematosaqua-
Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree

14 Übeyli ED. Multiclass support vector machines for diagnosis of erythmato-squamous diseases. Expert Systems with Applications. 2008; 35(4): 1733-40. doi: http://dx.doi.org/10.1016/j.eswa.2007.08.067.

15 Übeyli ED, Güler I. Automatic detection of erythmato-squamous diseases using adaptive neuro-fuzzy inference systems. Computers in Biology and Medicine. 2005; 35(5): 421-33. doi: http://dx.doi.org/10.1016/j.compbiomed.2004.03.003.

16 Danjuma K, Ososfan AO. Evaluation of Predictive Data Mining Algorithms in Erythmato-Squamous Disease Diagnosis. International Journal of Computer Science Issues. 2014; 11(6): 85-94.

17 Maimon O, Rokach L. Data mining and knowledge discovery handbook: Springer, 2005.

18 Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

19 Bache K, Lichman M. "(UCI) Machine Learning Repository". University of California, Irvine, School of Information and Computer Sciences, 2013.

20 Larose DT. Discovering knowledge in data: an introduction to data mining. Hoboken, N.J.: Wiley-Interscience, 2005.

21 Übeyli ED, Doğdu E. Automatic detection of erythmato-squamous diseases using k-means clustering. Journal of medical systems. 2010; 34(2): 179-84.

22 Soltany TS, Langarizadeh M, Shanbezadeh M. Automatic detection of erythmato-squamous diseases using K-Nearest Neighbor Algorithm. Iranian Journal of Medical Informatics. 2013; 2(2).