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

Keivan Maghooli¹, Mostafa Langarizadeh², Leila Shahmoradi³, Mahdi Habibi-koolae⁴,⁵, Mohamad Jebraeily⁶, and Hamid Bouraghi⁷

¹Biomedical Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran  
²Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran  
³Department of Health Information Management, Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran  
⁴Health Management and Social Development Research Center, Suleiman University of Medical Sciences, Gorgan, Iran  
⁵Department of Health Information Technology, School of Allied Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran  
⁶Department of Librarianship and Medical Information, Hamadan University of Medical Sciences & Health Services, Hamadan, Iran  
⁷Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

Corresponding author: Mahdi Habibi-koolae, Health Management and Social Development Research Center, 3rd Golbarg Alley, Golha Str., Gorgan, Iran. Phone: +98 17 32160330-1; E-mail: habib@goums.ac.ir

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. Olatunji 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 pursue 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.

Recurrent 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).

Table 1. Clinical and histopathological features of dermatology data set

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

Table 2. Clinical and histopathological features of dermatology data set and the classes

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 (1). 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. Ifter from Gazi University and H.A. Guvenir from Bilkent University (1, 7).

The dermatology data set contains 366 samples and 34 at-
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 CARD 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 tL and tR are left and right child node of t, respectively. PL and PR are the number of records at tL and tR 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).

 foresight

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 the diseases into correct or wrong classes.

As confusion matrix show, one of the cases of psoriasis classified as seborheic dermatitis by the model. For lichen planus, the model classified one case as seborheic dermatitis and another case as pityriasis. Also, four cases of pityri-
Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree

asis rosea classified as seborrheic dermatitis and two cases of chronic dermatitis and pityriasis rubra pilaris classified as psoriasis.

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.

REFERENCES

13 Abdi MJ, Giveki D. Automatic detection of erythematous-squa-
Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree


INSTRUCTIONS FOR THE AUTHORS

All papers need to be sent to: Editorial board of the journal Mat Soc Med, electronically on the web site http://scopemed.org/?jid=16. Every article sent to Mat Soc Med gets its number, and author(s) will be notified if their paper is accepted and what is the number of paper. Original paper could not contain more than 3,000 words. Review article more than 4,500 and Case report more than 1,500 words, including References.

Every correspondence will use that number. The paper has to be typed on a standard format (A4), leaving left margins to be at least 3 cm. All materials, including tables and references, have to be typed double-spaced, so that one page has no more than 2000 alphabetical characters (30 lines) and total number of used words must not to be more than 3,500. Presenting paper depends on its content, but usually it consists of a title page, summary, text references, legends for pictures and pictures, type your paper in MS Word and send it on a diskette or a CD-ROM, so that the editing of your paper will be easier.

Every article has to have a title page with a title of no more than 10 words: name(s), last and first of the author(s), name of the institution the author(s) belongs to, abstract with maximum of 45 letters (including space), footnote(s) with acknowledgments, name of the first author or another person with whom correspondence will be maintained.

The paper needs to contain structured summary, 200 words at the most. Summary needs to hold title, full name(s) and surname(s) of the author(s) and coauthor(s), work institution, and all essential facts of the work, introduction, formulation of problems, purpose of work, used methods, (with specific data, if possible) and basic facts. Summary must contain the review of underlined data, ideas and conclusions from text. Summary must have no quoted references. Four key words, at the most, need to be placed below the text.

Authentic papers contain these parts: introduction, goal, methods, results, discussion and conclusion. Introduction is brief and clear review of the problem. Methods are shown, so that interested reader is able to repeat described research. Known methods do not need to be identified, they are cited (referenced). If drugs are listed, their generic name is used, (brand name can be written in brackets). Results need to be shown clearly and logically, and their significance must be proven by statistical analysis. In discussion, results are interpreted and compared to the existing and previously published findings in the same field. Conclusions have to give an answer to author’s goals.

Quoting references must be on a scale, in which they are really used. Quoting most recent literature is recommended. Only published articles, (or articles accepted for publishing), can be used as references. Not published observations and personal notifications need to be in text in brackets. Showing references must be as how they appear in the text. References cited in tables or pictures are also numbered according to the quoting order. All references should be compiled at the end of the article in the Vancouver style or PubMed style (i.e. www.scopemed.org).

Tables have to be numbered and shown by their order, so they can be understood without having to read the paper. Every column needs to have a title, every measuring unit (SI) has to be clearly marked (i.e. preferably in footnotes below the table, in Arabic numbers or symbols). Pictures also have to be numbered as they appear in the text. Drawings need to be enclosed on a white or tracing paper, while black and white photos have to be printed on a radiant paper. Legends (e.g. next to pictures and photos), have to be written on a separate A4 format paper. All illustrations, (pictures, drawings, diagrams), have to be original, and on their backs contain, illustration number, first author’s last name, abbreviated title of the paper and picture at the top. It is appreciated, if author marks the place for the table or picture. Papers could not be submitted with included more of 4 Tables, Figures or Graphs. Every additional must be paid 20 euros each.

Use of abbreviations have to be reduced to a minimum. Conventional units can be used without their definitions. Supplement, if paper contains original contribution to a statistical method or author believes, without quoting original computer program, that paper’s value will be reduced. Editorial staff will consider possibility of publishing mathematics/statistic analysis in extension.

Any practice of plagiarism will not be tolerated regarding submitted articles. Non-identifiable quoted parts of the articles from other authors are known as plagiarism if it is not cited or referencing in appropriate places in the article. Advertent practice of plagiarism will abort reviewing process or article submission. Author(s) may suggest or exclude peer-re-viewers for their articles but Editorial Board has the right to reject their(s) opinions or suggestions according to copyright Assignment form signed by authors before reviewing process. Authors must respect guidelines and rules of ICMJE, WAME, COPE, EASE, linked on www. avicenapublisher.org.

All individuals listed as authors should qualify for authorship and should have participated sufficiently in the work to take public responsibility for appropriate portions of the content and follow the next conditions: a) substantial contributions to the conceptions and design, acquisition of data, or analysis and interpretation of data; b) drafting the article or revising it critically for important intellectual content; c) final approval of the version to be published (all co-authors must sign copyright Assignment form downloaded from www.avicenapublisher.org). All other contributors to the article’s subject who does not qualify for authorship should be listed in acknowledgement section, for all relevant information about authorship follow ICMJE guidelines.

All authors must make a formal statement at the time of submission indicating any potential conflict of interest that might constitute an embarrassment to any of the authors if it were not to be declared and were to emerge after publication. Such conflict of interest might include, but not limited to, share holding in or receipt of grant or consultancy free form a company whose product features in the submitted manuscript or which manufactures a competing product. All authors must submit a statement of conflict of interest to be published at the end of their article (conflict of Interest: non declared).

342 ORIGINAL PAPER / ACTA INFORM MED. 2016 OCT; 24(S): 338-342