ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

PREDICTION OF DERMATOLOGICAL CONDITION USING NAÏVE BAYESIAN CLASSIFICATION

Sistla Karthik*, Bhimana Sarath Rahul ** Mukesh Vibhudhi *, Tarun Tej K*****
*Scope, VIT University.
**NRI Institute of Technology.
***KL University.
****PVP Siddhartha Institute of Technology.
Email: bhimanasarathrahul@gmail.com

Received on: 18-02-2017 *Accepted on: 22-03-2017*

Abstract:

Abstract –The study of different types of skin diseases is known as dermatology. Dermatological diseases if neglected can cause death or terrible problems. It is difficult to predict dermatological diseases due to the many number of attributes required to identify a disease. We deal with the major 6 categories of skin diseases. The algorithm used is the Naïve Bayesian classification algorithm. Based on the parameters the algorithm gives us the probability of occurrence of a particular kind of dermatological disease along with the percentage. The model which is used for testing is taken from the dermatology dataset downloaded from the UCI repository site.

Keywords: Dermatology, Naïve Bayesian, skin, Clinical Data Mining

Introduction:

Healthcare and medical reports usually generate huge amounts of data that are difficult to analyze and study. So we make use of data mining which helps in easier analysis, identifying patterns and association rules among the datasets and their parameters which is very useful in the field of healthcare. Though there are numerous techniques in order to classify data we use the Naïve Bayesian technique to classify the types of dermatology diseases which is modelled on random probabilistic theory Dermatology is the branch of science that deals with the skin. Skin diseases are very difficult to diagnose as they are based upon a number of parameters and are hard to identify. These diseases if not diagnosed and treated at the right time cause death in some cases. We focus on the 6 common types of skin diseases namely chromic dermatitis, lichen planus, pityriasisrosea, pityriasisrubrapilaris, psoriasis, seboreic dermatitis. Based upon the given

parameters we create a classification tree for the data and classify the data for identification of the disease using the naïve Bayesian algorithm. This makes identification easier.

Literature Review:

Many authors have worked in the field of dermatology by using numerous data mining algorithms. In the paper byBruno Fernandes Chimieski1etal.(1) the author has made use of the Bayesian technique and logistics model tree technique. In the paper by Güvenir, H., Demiröz, G[2] the author has proposed an algorithm named VF15 and has managed to achieve an accuracy of 96%. In the paper presented by Guvenir, H., &Emeksiz, N. [3] the author proposes a system using naïve Bayes nearest neighbor and VF15. The author Nanni[4] has suggested using support vector machines with random subspaces and has attained admirable results .Polat and gunes[5] have used C 4.5 and hasattained accuracy of 96.25%.Whereas Ubeyli et.al[6][7] has used multiclass error correcting and k meansclustering methods and has attainedpleasing results.

In paper byDinesh K.Sharma, Hota H.S [8] authors have established hybrid model based on SVM and artificial neural network and obtained99.25%of accuracy, while in [9] another hybrid model is established on the basis of multilayer perceptron, Decision tree .In this study Bayes net with Best First search has been applied to the dermatology dataset downloaded from UCI repository site [10]. This study aims to categorize the dermatology diseases with best possible accuracy and reduced set of features.

Algorithm Used-Naive Bayes Classifier

The Naive Bayes Classifier is a very renowned algorithm due to its simplicity, computational efficiency and its amazingly decent results for existing practical problems. For instance numerous email clients like Mozilla Thunderbird or Microsoft Outlook use naive Bayesian classifiers in order to filter theannoying emails.

The model normally has the assumption that all the listed attributes are highly independent which a distant dream in reality is. This is why the name Naive is in existence. Even having violations of the simple conventions and the ingenuous design of the classifier, they are also appropriate for normal distributions, which are not uncommon in real-world problems. The Bayes classification is a supervised learning technique and also a statistical technique for classification. Probabilistic model is used to find out the uncertainty about the model by predicting the probabilities of the outcomes in a systematic way. In general, this naive Bayes classifier assumes that the presence of a particular feature

of a class is no way related to the presence of any other feature. For example, a person may be considered to be a lady if she is short, has long hair and has curves. Even if these features are not same for everyone and vary from one to other, a naïve Bayes classifier reflects all these stuffs to independently contribute to the probability that the person is a female.

This classifier has the following advantages:-

- Training and implementation is an easy task. It is versatile in nature. Suppose, if the naive Bayes conditional independence is valid then the classifier will perform faster when compared to other training methods.
- The naive Bayes classifier performs better even if the conditional independence is not valid
- The probability distribution of classifier gives very accurate prediction. If the accuracy is not up to mark then subsequently then ignore the prediction.
- Class imbalance can be compensated wherein one or more than one instances occur very rarely (1 /1000). This can lead to wrong and incorrect predictions which is called degenerate solution. In such situations balanced training set must be used.
- When the training data set is very big it works in a very accurate manner.
- The training time complexity of this classifier makes thelearning technique both time as well as storage efficient.

Bayes' Theorem: Probability (B given A) = Probability (A and B)/Probability(A)

In order to get the probability of B given A, the no of cases where A and B occur together is taken into consideration and is divided by the no of cases where A alone occurs without the occurrence of B. If we assume Y is any data tuple, Y is considered as evidence of a particular situation and assume H is any hypothesis. Now if Y belongs to a particular class C P(H|Y) is the probability, of H conditioned on Y.

The Bayes theorem is mathematically denoted by the following formula

$$P(H/Y)=\frac{P(Y/H)*P(H)}{P(Y)}$$

Why Naive Bayes Classifier?

Naïve Bayesian Classification plays a crucialrole in medical data mining due to its great accuracy. Naïve Bayesian classifier makes use of independent attributes, therefore we can use it in medical fields for prediction if the attributes are

independent of each other. Naïve Bayes also handles missing values efficiently which are quite common in medical/clinical data. One of the main advantage is that the error rate of the naïve Bayesian classifier is very low.

Tool Used:

Orange is a machine learning and data mining software which is written in python. The program is upheld and developed by the Bioinformatics Laboratory of the Faculty of Computer and Information Science at University of Ljubljana.

It is a component-based visual programmingsoftware.Components are called widgets and they range from visualizing simple data and selecting the required subsets from a given quantity of sets and preprocessing data for ease of access, to empirical evaluation of learning algorithms and predictive modeling.

Methodology and Obtained Results:

We have taken the dermatology data set used in this paper from the UC Irvine Machine Learning data repository. The data set is very detailed and classifies the diseases based on 34 attributes like erythema, scaling, family history, age and a variety of other factors. These attributes are used to predict the six major types of skin diseases namely chromic dermatitis, lichen planus, pityriasisrosea, pityriasisrubrapilaris, psoriasis, seboreic dermatitis. We make use of the orange tool and load the data set into the data table and visualized it in the form of a scattered plot. Now a data sample is taken into consideration with majority of the data being included. The classification tree is constructed using the classification tree viewer present in the orange tool.

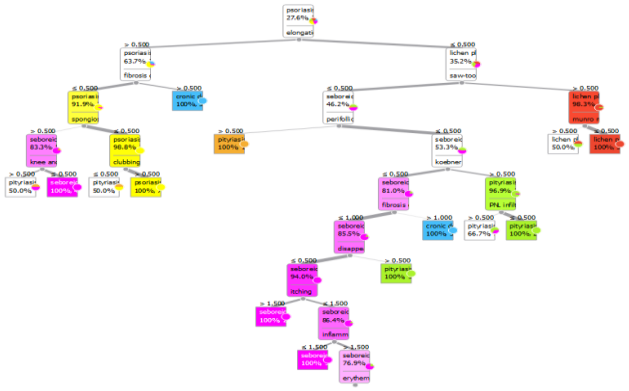

Fig1: The classification tree generated using the orange tool up to 9 levels.

We take /different algorithms like Naïve Bayesian Classifier, Support Vector Machine and Random Forest Classification in order to find out the accuracy of each method. The accuracy if each method is measured using the classification accuracy (CA) feature and the best method is taken into consideration.

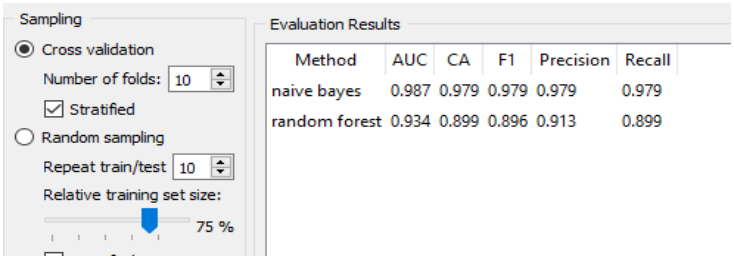

Fig2: Shows the Classification Accuracy of Naïve Bayesian and Random Forest Classification.

From the figure we can clearly observe that the CA ofNaïve Bayesian is higher than that of Random forest Classification. So we make the predictions using the Naïve Bayesian Classification Method. The Naive Bayesian predicts the probability of each type of disease by taking into consideration the various factors and finally displays the type of skin condition the patient is suffering from with high levels of accuracy compared to other algorithms.

Fig3: The scale in which each color denotes a specific disease.

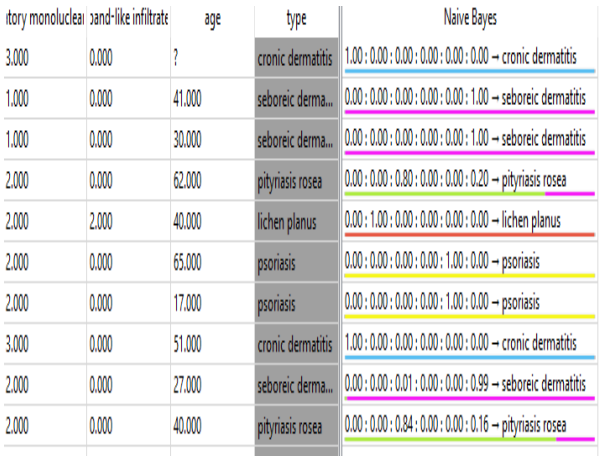

Fig4: Denotes the Predictions for a limited amount of data.

Prospectus and Future Work:

Clinical data mining is one of the widely researched areas nowadays and there is tremendous scope for updates and improvements in this regard. More techniques that deal with large amounts of health related data and techniques which are effectively able to handle missing data values should be encouraged as medical data usually has a large amount of

missing values. The prediction of diseases shaving a smaller no of attributes to classify should be made more efficient by developing tools for mining of smaller data and not compromising on the accuracy if classification

Conclusion:

Predicting the six types of dermatological diseases with high accuracy have been made possible with the help of Naïve Bayesian Classifier by analyzing the various (34) attributes of the patients. The model is an effective tool to analyze the type of condition given a large no of attributes for prediction. This is best possible effective model for the prediction of dermatological diseases. The high level of accuracy is another primary reason for choosing this method.We can extend this work with other data mining techniques and other medical measurements besides the above list. We can also predict diseases other than dermatological diseases which may lead to skin caner

References:

1. DivyaTomar and Sonali Agarwal, " A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-technology*, Vol.5, No.5, pp. 241 – 266, 2013.
2. Jiawei Han and MichelineKamber, *Data Mining: Concepts and Techniques*,2nded.,Morgan Kaufmann Publishers., An Imprint of Elsevier, 2006.
3. E.Barati, M. Saraee, A.Mohammadi, N. Adibi and M.R. Ahamadzadeh,"A survey on utilization of data mining approaches for dermatological (skin) diseases prediction," *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI):* March Edition, pp.1-11, 2011
4. G.Subhalakshmi, K.Ramesh and M. Chinna Rao, "Decision support in heart disease prediction system using naive bayes," *Indian Journal of Computer Science and Engineering (IJCSE),* Vol.2, No.2, pp.170-176, April-May 2011.
5. Kenneth Revett, Florin Gorunescu, Abdel – Badeesh Salem and El-Sayed El-Dahshan, "Evaluation of the feature space of erythematosquamous dataset using rough sets," *Annals of University of Craiova, Math. Comp. Sci. Ser.* Vol. 36(2), pp. 123 – 130, 2009.
6. DariuszMatyja, "Application of data mining algorithms to analysis of medical data," Master Software Engineering thesis, Blekinge Institute of Technology, Ronneby, Sweden, Aug. 2007.
7. Applications, 18(1) 43–49,2000.[8] Han, J., Kamber, M., and Pei, J.,.Data mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, San Francisco, CA. USA,2011.

8. Dinesh K.Sharma, Hota H.S. ,"Data Mining techniques for prediction of different categories of dermatology diseases", Academy of information and management science journal ,Vol. 16
9. (2), 103-115, 2013.
10. Krzysztof Cios, Witoldpedryeyz, Roman Swiniarski."Data Mining Methods for Knowledge Discovery", Kluwar Academic Publishers,1998.
11. Nanni, L. "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases.", Neurocomputing,69, 842–845,2006
12. Polat, K., &Güneş, S.. "A novel hybrid intelligent method based on C4.5 decision tree classifier and one against-all approach for multiclass classification problems". Expert Systems with Applications,36(2) 1587-1592,2009
13. Übeyli, E., &Dogdu, E. "Automatic detection of erythematosquamousdiseases using k-means clustering.", Journal of Medical Systems, 34, 179-184,2010
14. Übeyli, E., &Guler, I," Automatic detection of erythematosquamousdiseases using adaptive neurofuzzyinference system.",Computer in biology and medicine, 35, 421-433,2011
15. UCI (2013). Web source:http://archive.ics.uci.edu/ml/datasets.html,last accessed on Dec 2013
16. Guvenir, H., &Emeksiz, N. "An expert system for the differential diagnosis of erythe-matosquamous diseases.", Expert Systems with Applications, 18(1) 43–49,2000
17. Bruno Fernandes Chimieski1, Rubem Dutra RibeiroFagundes," Association and Classification Data Mining Algorithms Comparison over Medical Datasets", J. Health Inform. Abril-Junho*; 5(2): 44-5,2013.*
18. Güvenir, H., Demiröz, G., &Ilter, N. "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals". Artificial Intelligence in Medicine, 13(3) 147-165,1998
19. Elsayad, A. M. "Diagnosis of erythemato-squamous diseases using ensemble of data mining methods*"*, ICGST-BIME Journal, 10(1), 13- 23, 2010.