ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Classification of Dermatology Diseases through Bayes net and Best First Search

MadhuraRambhajani[1], Wyomesh Deepanker[2], Neelam Pathak[3]

M-tech Scholar, Technocrats Institute of Technology (Excellence), Bhopal, India[1]

Assistant Professor, Technocrats Institute of Technology (Excellence), Bhopal, India[2,3]

Abstract: In this world near about 1/7th of total world population suffer from some sort of skin disorder. The study of different types of skin disease or disorder is known as Dermatology. There are six different categories of skin diseases which shares somewhat same features. So for the classification of these diseases bayes net a Bayesian technique along with feature selection has been used in this study. Performance of all model are calculated using some measures like accuracy, sensitivity, and specificity. Model is tested from dermatology dataset downloaded from UCI repository site. After eliminating 20 features from dataset 99.31% of accuracy is achieved.

Keywords: Dermatology, Bayesian Technique, Feature selection, Best First Search**.**

I. INTRODUCTION

The medical data acquired electronically from patients has been drastically increasing day by day. Medical data include patient's records, employee record, pharmacy, laboratory etc grow in scope and capacity. This data should be handled properly so that it can be mined to provide relative information. Increased complexity in medical data has led to the emergence of development of Decision support system(DSS) for medical applications.[3] A decision Support System is an information system that collect ,organize and analyzes data for decision making activates[3]..DSS have been used in diseases diagnosis; prescription and other hospital management activities.DSS improve the efficiency of care by reducing the amount of time that a doctor spends on administrator tasks.

Information technology has provided a variety of computational machine learning algorithms such as Bayesian techniques, support vector machine etc which has been applied for detection of various complex diseases. In this study bayes net the Bayesian technique with Best first search feature selection has been used in order to classify the dermatology diseases algorithm to its corresponding types. Bayes net [10] is a Bayesian techniques which is based on random probabilistic theory of variables, where as Best first search [10] is a feature selection method based on greedy hill climbing method. [5].Dermatology is the branch of medical science that deals with skin diseases that is very difficult to diagnose that can be lead to harmful diseases like cancer [9]. The diseases in this group are psoriasis, exoreic dermatitis, lichen planes, pityriasis rosea, chronic dermatitis and pityriasis rubra pilaris [3] [4] [15] .They all share almost the same clinical feature of erythema and scaling with very little difference.

There are various authors who have contributed and have used various data mining algorithms for the diagnosis of dermatology diseases .In paper [3] Bayesian technique and LMT (Logistics Model Tree) model and have achieved confidence level greater than 100. H.Altay Guvenir et.al [6] has proposed a new classification algorithm VFI5 that is voting feature interval and has achieved 96.25% of accuracy. In paper [7] authors have extended their work and presented an expert system using naïve bayes nearest neighbor and VFI5. Nanni[11] has used support vector machine with random subspace and has achieved good results.Polat and gunes[12] have used C 4.5 and has achieved accuracy of 96.25%.Whereas Ubeyli et.al [13][14] has used multiclass error correcting and k means clustering methods and has achieved satisfactory results. In paper [9] authors have developed hybrid model based on SVM and artificial neural network and obtained 99.25%of accuracy, while in [4] another hybrid model is developed based multilayer perceptron, Decision tree and LDA.

In this study bayes net with Best First search has been applied to the dermatology dataset downloaded from UCI repository site [15]. This study aims to classify the dermatology diseases with highest accuracy and reduced set of features.

II MATERIALS AND METHODS

A) Materials

The material which is required for this study is the dermatology dataset .The dermatology dataset is taken from the university of California at Irvine (UCI) machine learning repository [9][15] to demonstrate the technique. Every instance of dataset is classified into six different categories.

The dataset contain 35 attributes in which 34 features are considered as input and the 35th feature is considered as target (class).There are 365 instance in the dataset, 112 instance belong to the psoriasis class, 60 instances belong to the seboreic dermatitis class, 72 instances belong to the lichen planus class, 49 belong to chronic dermatitis class and instances belong to the pityriasis rosea class, 52 instance 20 instance belong to pityriasis rubra pilaris class [9] [3].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

B) Methods

Bayesian net [10] [3] is a statistical processing based on bayes decision theory and is a fundamental technique for pattern recognition and classification. The Bayesian approach assumes that pattern possesses random characteristics and they are generated in a random way by some natural phenomena and process [10].

It is a graphical model that encodes probabilistic relationships among variable of interest. The natural choice for dealing with random and uncertain pattern is to use statistical technique based on probabilistic characteristics of data. The Bayesian method is based on the assumption that the classification of patterns is expressed in probabilistic terms.

It assumes that the statistical characteristics of random patterns are expressed as known probability values describing the random nature of pattern and their features. These probabilistic characteristics mostly concern a priori probability and conditional probability density of pattern of class.

The bayes decision theory provides a framework for handling the required probability descriptors of pattern processing problem. It provides statistical methods for classifying patterns into class based on probabilities of patterns and their features.

Feature selection [10] is a process of finding the best feature subset from original set of features, according to some defined feature selection criterion, without feature construction or transformation.

Feature selection (also known as subset selection) is a process commonly used in machine learning; where in a subset of the features available from the data are selected for application of a learning algorithm.

The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions**..**

There are various methods of feature selection , in this study Best First search has been used.. Best first search[5] is a general heuristic based search function. In best first search, in the graph of problem representation, one evolution function is attached with every node.

The value of evolution function is based on cost or distance of current node from goal node. The decision of which node is expanded is depend on the value of this evolution function

III MODEL IMPLEMENTATION

Model implementation process composed of three steps given in fig:1.(i) Data Preparation (ii) Model development (ii)Model validation .In Data preparation data is prepared to fed into the model after that attribute selection is performed so that relevant features are extracted while irrelevant are removed and finally data is partitioned into two mutually exclusive dataset I.e. training dataset and testing dataset in the ratio of 60:40.

In model development and model validation steps, a classification model with desired accuracy is first developed then performance is measured using error measures like accuracy, sensitivity, and specificity.

Classification model is a way through which one can describe data into their respective classes. In this procedure, a model or classifier is constructed to predict categorical labels

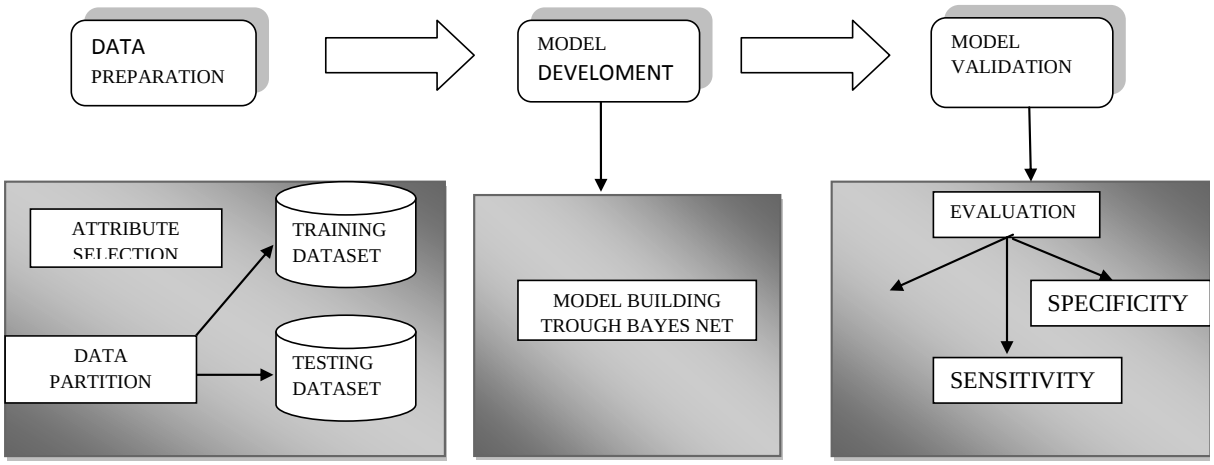

Fig I : Model Framework

DOI 10.17148/IJARCCE.2015.4526

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

IV. EVALUATION METRICES

The performance of the individual models are calculated using various statistical measures; accuracy, sensitivity, specificity F-measure TP rate, Precision [3][9][[4] etc. These measures are defined using true positive, true negative, false positive and false negative[3]. A true positive [4] [9] decision occurs when the positive prediction of the system matches with a positive prediction of a dermatologist. A true negative [4][9] decision occurs when both the system and the physician predict the absence of positive predictions. False positive [9] occurs when a system labels a healthy case; a negative one as a positive case. Finally, false negative [4][9] occurs when the system labels a Positive case as negative. Table 1 represents a matrix showing cases of TP,TN,FP,and FN which is known as confusion matrix[4][9].

Following are some of the performance measurements which can be calculated using above table:

Accuracy: Accuracy [3] can be defined as the no of instances classified correctly by the total number of cases

Accuracy = (TP + TN) / (P + N)……..... (1)

Sensitivity: Sensitivity [3] measures the possibility of positive classification of instances i.e. TP to the sum of TP and FN

Sensitivity = TP / (TP + FN) …… (2)

Specificity: Specificity [3] measures the possibility of negative classification of instances i.e. TN to the sum of TN and FP

Specificity= TN / (TN + FP)…….. (3)

F-measure: F measure [3] produces higher output when precision and recalls both are balance and is important.

Precision: Precision [3] is the proportion of the predicted pages that is

TP/FP+TP…….. (4)

Recall [3] is the proportion of the relevant pages that were correctly classified i.e.

TP/FN+TN……. (5)

F measure [3] is derived from recall and precision I.e.
2*Recall * precision/Recall + Precision…….. (6)

Table I: STRUCTURE OF CONFUSION MATRIS FOR TWO CLASS PROBLEMS

Actual Vs Predicted	Positive (P)	Negative(N)
Positive(P)	True Positive(TP)	False Negative(FN)
Negative(N)	False Positive(FP)	True Negative(TN)

V. RESULTS AND DISCUSSION

In these section results of classification and of different dermatology diseases has been reported. In classification first after reduction of features and data partition we have obtained best results with 15 features of erythemato squamous dataset , a confusion matrix is obtained which is used to identify the performance of a classifier . Here Table 2 presents confusion matrix for testing dataset which specifies no of instances correctly classified and misclassified. And all performance measurements are calculated using above mentioned

TABLE II Confusion matrix of classifier for testing dataset

	Predicted Class					
Actual class	**psoriasis**	**seboreic Dermatitis**	**lichen Planus**	**pityriasis Rosea**	**chronic Dermatitis**	**pityriasis Rubra Pilaris**
psoriasis	44	0	0	0	0	0
seboreic dermatitis	0	26	0	1	0	0
lichen Planus	0	0	26	0	0	0
pityriasis Rosea	0	0	0	23	0	0
chronic Dermatitis	0	0	0	0	17	0
pityriais Rubra Pilaris	0	0	0	0	0	9

In the above table II each cell in the table contains the number of instances in a particular class. The predictions are compared with the actual classes to determine true positive, true negative false positive ,false negative.

For example the number of classified instances for seboreic dermatitis is 26 and the number of instances misclassified is 1.With the help of above table and using equation 1, 2 3 ,4, 5 and 6 the performance measurement of each class and average value of model has been specified in table III

TABLE III Value of different statistical measures for different classes

	Performance measurement								
Class	***Sensitivity***	***Specificity***	***Accuracy***	***TP rate***	***FP rate***	***F-measure***	***Precision***	***Recall***	***ROC***
Psoriasis	100	100	100	1	0	1	1	1	1
Seboreic dermatitis	96.30	100	96.29	0.963	0	0.981	1	0.963	1
Lichen Planus	100	100	100	1	0	1	1	1	1
Pityriasis Rosea	100	99.20	100	1	0.008	0.979	0.958	1	1
Chronic Dermatitis	100	100	100	1	0	1	1	1	1
Pityriais Rubra Pilaris	100	100	100	1	0	1	1	1	1
Average value	99.3	99.9	99.31	0.993	0.001	0.993	0.993	0.993	1

VI. CONCLUSION

There is great attention of research with single technique but the results are more promising and is time taking with large number of features. In this study an integrated technique of a Bayesian technique along with a feature selection has been proposed. In this bayes net a Bayesian technique along with best first search feature selection has

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

been applied to the dermatology dataset and after eliminating15 features from the dataset accuracy of 99.31 of accuracy is obtained.

REFRENCES

[1] Arun K. Pujari, Data mining techniques, 4th edition, Universities Press (India) Private Limited,2001.
[2] Barati, E., Saraee, M., Mohammadi, A., Adibi, N, & Ahamadzadeh, M.R. ." A survey on utilization of data mining approaches for dermatology (Skin) disease prediction," Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI), March Edition, 1-11,2011.
[3] Bruno Fernandes Chimieski1, Rubem Dutra Ribeiro Fagundes," Association and Classification Data Mining Algorithms Comparison over Medical Datasets", J. Health Inform. Abril-Junho*; 5(2): 44-5,2013.*
[4] Elsayad, A. M. "Diagnosis of erythemato-squamous diseases using ensemble of data mining methods",ICGST-BIME Journal, 10(1), 13-23,2010.
[5] Elaine Rich ,Kevin Knight,Artificial Intelligence, McGraw-Hill, 01-Jan-1991
[6] 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.
[7] Guvenir, H., & Emeksiz, N. "An expert system for the differential diagnosis of erythemato-squamous diseases.", Expert Systems with 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.
[9] 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 (2), 103-115, 2013.
[10] Krzysztof Cios,Witold pedryeyz,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 oneagainst-all approach for multi-class classification problems". Expert Systems with Applications, 36(2) 1587-1592,2009
[13] Übeyli, E., & Dogdu, E. "Automatic detection of erythemato-squamous diseases using k-means clustering.",Journal of Medical Systems, 34, 179-184,2010
[14] Übeyli, E., & Guler, I," Automatic detection of erythemato-squamous diseases using adaptive neuro fuzzyinference 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.

BIOGRAPHIES

Madhura Rambhajani is a M-tech scholar in Department of Information technology, Technocrats Institute of Technology,(Excellence) Bhopal, India. She received B.E in 2012 from Guru Ghasidas Vishwavidyalaya Bilaspur (C.G).

Wyomesh Deepanker currently working as Assistant Professor in Technocrats Institute of Technology, (excellence)Bhopal, India. His area of specializations is soft computing, Data Mining, Bio inspired Computing, Image Processing.

Neelam Pathak currently working as Assistant Professor in Technocrats Institute of Technology(excellence), Bhopal, India. Her area of Specialization is Computer network and soft computing.