AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Diagnosis of Erythemato-Squamous Disease Using Data Mining Techniques.

[1]Geetha BK and [2]Chandra J

[1]*Geetha BK, PG Scholar, Department of Computer Science, Christ University, Bangalore, Karnataka, India,*
[2]*Chandra J, Associate Professor, Department of Computer Science, Christ University, Bangalore, Karnataka, India,*

Address For Correspondence:
Geetha BK, Geetha BK, PG Scholar, Department of Computer Science, Christ University, Bangalore, Karnataka, India,

ARTICLE INFO

Article history:
Received 18 January 2017
Accepted 28 March 2017
Available online 15 April 2017

Keywords:
Erythemato-Squamous disease, Classification algorithms, Bagging, NavieBayes, BayesNet, Logisti Boost, Performance evaluation.

ABSTRACT

Background: The purpose of the paper is to analyze Erythemato-Squamous disease using data mining technique. Identification of different Erythemato-Squamous is a troublesome issue in dermatology. This is the case with Lichen planus, Pityriasis rosea, Pityriasis rubra pilaris, Psoriasis, Seboreic dermatitis and Cronic dermatitis. Since, all side-effects share the clinical component of erythema and scaling with extremely edge contrast at the initial stage. Clinical evaluation is done for 12 features at the beginning stage and the skin samples are considered for the evaluation of 22 histopathological features. **Objective:** The motivation behind the research work is to demonstrate the consequence of data mining algorithms which have only clinical feature as input in helping the physician to categorize six different types of Erythemato-Squamous disease and also to achieve accurate prediction connected on clinical dataset of Erythemato-Squamous disease from UCI (University Of California, Irvine) machine repository. **Results:** The current research focused on using Bagging, Bayesian network, Support Vector Machine (SVM), Logic boost, Multilayer Perceptron (MLP), Random Forest, J48 data mining classifier for the prediction of the disease.**Conclusion:** With the help of experimental evaluation, it is observedthat Bayesian Network and Logic Boost algorithms were able to classify the diseases with the accuracy of 83.6% and 84.5. Hence Bayesian Network and Logic Boost algorithms were giving better accuracy than other data mining classification.

INTRODUCTION

The process of discovering interesting and useful patterns and relationship in large volume of data is known as data mining. In real world applications, a data mining process can be phased as data understandings, data preparation, modeling, evaluation and deployment. Data Mining collectively referred as Knowledge Discovery in Database (KDD) (X. Li *et al.*, 2009). Data mining in healthcare holds great potential to use data systematically and identify best practices to improves care and reduce cost (Peng Y *et al.*, 2010). The use of computer technology plays an important role in prediction of surgical procedures, medical tests, medications and discovery of relationship between pathological data and clinical data, finding malignant tumors in different organs of body (A.Oztekin *et al.*, 2009; K. Imran *et al.*, 2009; W. Moudani, 2013). Hence the application of data mining for healthcare and biomedical is increasing day by day (J.M.Renders and T.Simonart, 2009)**.**

Skin is the biggest organ of the body which covers an aggregate region around 20 square feet in normal grown-up. It is the main line of protection against microorganisms and harm and frequently mirrors the general strength of the body. The most common disease comes under dermatology are Acne, Dermatitis, Fungal infection, Hair disorders, Nail problem, Psoriasis, Rosacea, Skin cancer, Shingles, Vitiligo, Warts. In this paper we are discussing about disease like Erythemato-Squamous being a very common to everyone, factors like

ToCite ThisArticle: Geetha BK and Chandra J., Diagnosis of Erythemato-Squamous Disease Using Data Mining Techniques. ***Aust. J. Basic & Appl. Sci.,*** *11(5):45-51, 2017*

drugs, microbes, exposed to ultraviolet radiation in sunlight causes this disease (A. Kampourakia *et al.,* 2013). In dermatology, the contradiction of different Erythemato-Squamous infection is a genuine issue science they all share the clinical elements of rednessand dry cracked skin as common symptom with amazingly slight differentiations. Lichen planus, pityriasis rosea, pityriasis rubra pilaris, psoriasis, seboreic dermatitis and cronic dermatitis are six major diseases which come under the Erythemato-Squamous. Ordinarily a biopsy is principal for the examination (Bekir Karlik and Gunes Harman, 2013). Classification techniques are widely used in almost all areas nowadays because of their knowledge discovery nature, It acts as intelligent decision maker as well as predictor tool. Hence they are used in education sector to analyze and predict the student‟s performance (Parneet Kaur *et al.,*2015), and to determine the online review‟s quality by using hidden topics distribution information (Hoan Tran Quoc *et al.,*2015), the kernel frameworks is used in big data analysis, offline learning, distributed database, online learning and its prediction (Yuichi Motai, 2012). Ensemble learning is machine learning process in which multiple classifiers are trained to solve the particular problem like with the insurance dataset from UCI repository a prediction model is built for know best policy investment. The AdaBoost and multiclassifier SVM ensemble produces the great accuracy to predict best investment policy (Chandra J and Siji T. Mathew 2012).The utilization of expert system as a mean of directing medical diagnosis and suggesting effective treatment has been a highly active from recent years. The algorithms like SVM, Neural Network, Bayesian Network, J48, Random Forest, Bagging, Logistic Boost etc., has been proved as a powerful promising tools for predicting and diagnosis disease (Nicholas I *et al.,* 2009; Mythili T *et al.,* 2013; Guosheng Wang, 2008).

MATERIALS AND METHODS

2.1 Existing Methods:

Latha Prathiban and R.Subramanian (2009) have proposed a Coactive Neuro-Fuzzy Inference System (CANFIS) model to detect the Erythemato-Squamous disease. The CANFIS classifier model is intended by aggregating the neural network adaptive capabilities and fuzzy logic qualitative approach which is then integrated with genetic algorithm. Records of 34 features of patients are diagnosed for six Eythemato-Squamous disease indications. CANFIS model has great impact in detecting the presence of disease.

Hatice cataloluk and Metin Kesler (2012) built a data mining tool to prognosis the Erythemato-Squamous disease by comparing the weighted K-NN with basic K-NN algorithms on UCI dermatology data set. Finally weighted K-NN algorithm gives the best result than basic K-NN. Similarly, comparison between the distances criteria is made between Euclidean distance is more fruitful than the Manhattan distance measure. In both weighted k-NN gives more accuracy than basic K-NN.

Avik Basu *et al.,* 2015 have made a comparative study on different kernel methods in SVM.SVM being a widely used machine learning and pattern recognizing methodology which can be used to both non-linear and linear classification. The three fold validation method is used to choose the optimal value of the parameters. By testing different kernel methods, it is found that linear kernel method gives the best accuracy over the dataset taken from UCI repository.

M.Shamsul Arifin *et al.,* 2013 have taken 704 skin images of total 2055 diseased areas with high resolution camera for six different Eythemato-Squamous diseases. The dataset is collected from ahospital in Bangladesh. Basically the system is made to work in two aspects one is for identify the diseased skin with the help of image processing, k-mean cluster and color gradient techniques and other is to detect the disease using feed-forward back-propagation Artificial Neural Network classification method. They have mentioned that the system exhibits great accuracy in detecting the diseased skin and also to identifying the diseases.

Bekir Karlik and Gunes Harman 2013 have built Computer Aided Software for the classification of Erythemato-Squamous disease. The supervised back propagation algorithm is used to train the network and the research work also makes a comparative study over different methods from other literature.

Chandra J *et al.,* 2013 have shown that SVM is the best methodology in diagnosing the different cervical cancer stages. The method helps to assess the impact short term and long term carcinomaS.Sasikala *et al.,* 2013 have proposed a improved normalized point wise mutual information (“INPMI”) model which helps in feature selection, classification accuracy and computational time. The model is coupled with Sequential Forward Search (SFS) to find the best feature selection processes. The experiment is conducted on UCI repository dermatology dataset. INPMI-SVM model with SFS has given the highest accuracy than INPMI-NB model with SFS and INPMI model J48 and also with SFS. Similarly, The INPMI also does a great work when it is experimented on World Aircraft dataset.

Davar Giveki *et al.,* 2013 have introduced a diagnostic model based on Catfish binary particle swarm optimization (Catfish BPSO), Kernelized support vector machine (KSVM) and association rule (AR) I.e.AR-CatfishBPSO-KSVM model as the feature selection technique to analyze Erythemato-Squamous disease.

Australian Journal of Basic and Applied Sciences, 11(5) April 2017, Pages: 45-51

2.2Methodology:

2.2.1 Boosting:

The AdaBoost algorithm is proposed by Yoav Freund(1990) and Robert Schapire(1995) in the computational leaming theory litrarure, it is a standout amongst the most vital gathering techniques, since it has strong theoretical base, accurate prediction with absence of complication and effective applications. Boosting works by constantly applying a classification algorithm to reweighted form of the training data and then by applying major voting policy a sufficient classifier is generated, Which is nothing but it is a method of consolidating the working of many classifiers to deliver an intense one (Yoav Freund and Robert E. Schapire, 1996). The A Bayesian network is a directed acyclic graph or DAG with a conditional probability distribution among variables of interest based on their probabilistic relation (Chen-Fu Chien *et al.,* 2002). Each node pictures a variable of the ranges over a definite set of domain and will have a connection with its parent‟s node. Each directed arc signifies the relationship of its variables and the conditional probabilities with its parent‟s node represents the degree of relationship of variables (Sho-Zhong Chang Hong yu *et al.,* 2003). The generation of Bayesian Network can divided into two functions, Structure learning, which deals with network topology of Bayesian Network from the data set and Parameter learning, which consists of calculating numerical parameters for the structures. Using improved EM Algorithm, the parameter learning has been more advantages than the standard EM algorithm (Shon-Zhong Zhang *et al.,* 2004).Bayesian

AdaBoost algorithm is combined with many other classifiers to get better result, like the combination of AdaBoost and Neural Network is used to classify the stored product insects, where a set of images of stored product insects are taken and 25 features are extracted from them. This features are used as the input of the AdaBoost-Back Propagation neural network with 20 hidden layer and 3 output nodes. The experimental result shows the better result with AdaBoost-BP neural network as compared with standard Back-Propagation system (Hongmei Zhang *et al.,* 2008).

2.2.2 BayesNet:

network architecture is used for handwriting recognition system mainly characters, writing degradation and also solving the cuts and missing data (Khlifia Jayech *et al.,* 2012) and the algorithm is also applied in labeling of documents. Logical labels of text blocks are compared with physical text components of content found in periodicals and magazines. The experiment results, Bayesian Network is sufficient for logical labeling in document (Souad SOuafi-Bensafi *et al.,* 2002).

2.2.3 Naive Bayes:

Naive Bayes is one of the most effective data mining classifier. This simple probabilistic classifier based on applying Bayes theorem works easily with huge datasets. It is very easy to build without any complicated iterative parameter estimation methods.

$$\mathrm{P(c/s)} = \frac{\mathrm{P(s/c)P(c)}}{\mathrm{P(s)}} \quad (1)$$

In equation 1, P is the probability that a set of symptoms “s‟ belongs to a particular class “c‟. Naïve Bayes algorithm is used in text classification as spam email detector. Initially all mails are preprocessed leaving the main body. With the tokenizer all stop words are deleted from the list. The vocabulary table in the model is used to generate the word map (Haiyi Zhang and Di Li, 2008).

2.2.4 Bagging:

Bagging is one of the most essential data mining method which is also called as bootstrap aggregation introduced by Breiman (Breiman *et al.,*1996). In bagging, each prediction is considered as add-on weightage to the classifier. Hence the classifier with maximum weight is considered for final classification. Good classifiers are transformed in optimal one in this method. The best subset can be expected always by best bagged predictor. Hence the bagging classifier is used in breast cancer prediction to differentiate between cancer and non-cancerPatient from 286 dataset of city hospital (Hemant Palivela *et al*). The bagging algorithm is also used in 3D object learning with Shape Spectrum Descriptor (SSD) approach to characterize the shape of the surface. Bagging is used with Random forest and Decision stamp with 10, 20 and 50 iterations. Both the algorithms work well with bagging in 3D object recognition (Omar HEROUANE *et al.,*2015).

3. Experimental Result:

3.1Dataset:

The Dermatology dataset from UCI repository has chosen to demine different the type of Erythemato-Squamous disease. The dataset consist of totally 34 attributes of dermatological condition which includes 22 histopathological and 12 clinical attributes from366 patients‟ record. Which specifically mentions about the six major Erythemato-Squamous diseases namely Lichen planus, Pityriasis rosea, Pityriasis rubra pilaris, Psoriasis,

Seboreic dermatitis andCronicdermatitisunder.

Table 3.1.1 The Erythemato-Squamous Disease Dataset.

Clinical features (values: 0,1,2,3)	Histopathological features (values: 0,1,2,3)
Scalp involvement	Munro microabcess
Koebner phenomenon	Eosinophils in the infiltrate
Oral mucosal involvement	Spongiosis
Polygonal papules	Fibrosis of the papillary dermis
Definite borders	Exocytosis
Follicular papules	Acanthosis
Age (linear)	Hyperkeratosis
Knee, elbow involvement	Perifollicular parakeratosis
Scaling	Clubbing of the rete ridges
Family history(0,1)	Elongation of the rete ridges
Erythema	Thinning of the suprapapillary epidermis
	Spongiform pustule
	Melanin inconsistence
	Focal hypergranulosis
	Disappearance of the granular layer
	Vacuolization and damage of basal layer
	PNL infiltrate
	Saw-tooth appearance of retes
	Follicular horn plug
	Parakeratosis
	Inflammatory monoluclear infiltrate
	Band-like infiltrate

The table 3.1.1 shows the Erythemato-Squamous disease dataset from UCI repository. In the dataset, family history feature has the value 1 if the disease noticed in family, and otherwise 0. Age of patients is represented by age feature. Every other feature has graded on a scale of 0 (zero possibility) to 3(largeamount of possibility), while intermediate level is represented by the values 1 and 2. The current research work has not histopathological data to construct the proposed prediction, since the study is to prediction of disease based on clinical data set.

RESULT AND DISCUSSION

Table 3.2.1: The performance evaluation on different classifiers using Erythemato-Squamous Data.

Classifier	Classification result					
	correctly classified	incorrectly classified	kappa statistic	mean absolute error	root mean squared error	Accuracy
RandomForest	77	23	0.622	0.172	0.280	70
J48	78	32	0.631	0.120	0.257	70.9
SVM	52	21	0.631	0.095	0.3097	71.2
MLP	55	18	0.686	0.087	0.2556	75.3
SimpleLogistic	86	24	0.723	0.089	0.2251	78.1
Lazy.IBk	86	24	0.726	0.100	0.2612	78.1
SMO	87	23	0.737	0.228	0.3202	79
Bagging	89	21	0.757	0.1152	0.2306	80.9
NavieBayes	90	20	0.773	0.088	0.2057	81.8
BayesNet	92	18	0.795	0.084	0.2099	83.6
LogitBoost	93	17	0.8056	0.0919	0.2011	84.5
NavieBayes	90	20	0.773	0.088	0.2057	81.8

The table 3.2.2 depicts the performance evaluation on different data mining techniques like SVM, MLP, Navie Bayes, and Logistic Boost etc on Erythemato-Squamous disease dataset. Although the number offeatures is reduced to 11 from 34 features, the classification is still successful high as seen in the table.

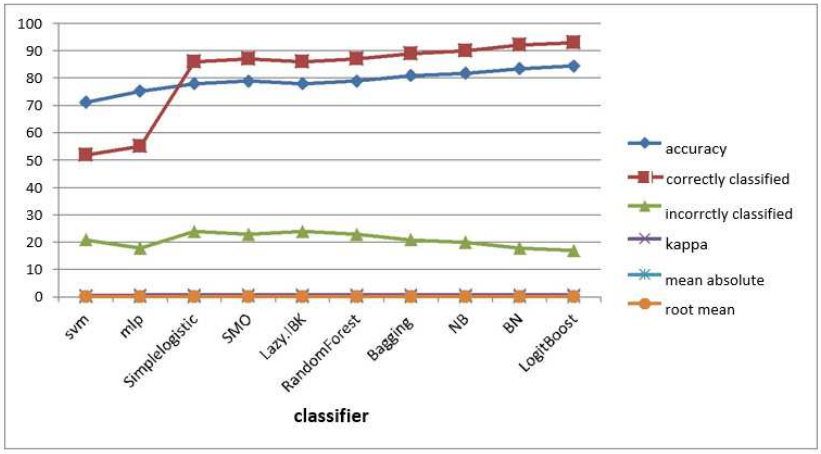

Fig. 3.2.2: Performance comparison on various data mining classifier using Erythemato-Squamous data.

The fig. 3.2.2 shows the graphical description of different classification methodologies and their accuracy level on dermatology dataset with 366 instances of Erythemato-Squamous disease. Thegraph clearly shows the LogitBoost, BayesNet, NavieBayes and Bagging algorithm performs well with the accuracy rate of 84.5%, 83.6%, 81.8% and 80.9%.

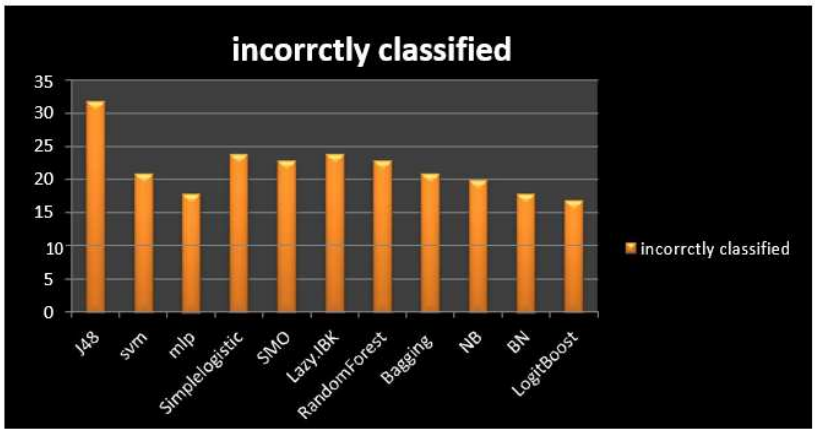

Fig. 3.2.3: Error rate comparison of different classifiers on Dermatology data

Fig 3.2.3 maps the error rate of the different algorithm. In the graph different classifiers are in the x-axis and number of instances is in y-axis. From the graph J48 algorithms has highest error rate with 32 incorrectly classified instances where as LogitBoost has very low error rate of 17 incorrectly classified instances.

Conclusion:

Dermatology is an area concerned with the health of skin, hairs, nails and mucous membranes. The dermatology data from UCI Repository states that diseases are very similar in terms of both clinical and histopathological features. Hence, the identification of different Erythemato-Squamous disease is a tedious work. The proposed work is concerned to medical diagnosis of Lichen planus, Pityriasis rosea, Pityriasis rubra pilaris, Psoriasis, Seboreic dermatitis and Cronic dermatitis diseases which comes under Erythemato-Squamous using only clinical attributes. The different data mining algorithms are used in the experiment. The Bayesian Network and logitBoost gives the highest accuracy result of 83.6% and 84.5%. Hence these two algorithms can be said better algorithm for further implementation with respect toonly with clinical feature of the Erythemato-Squamous disease.

ACKNOWLEDGMENT

The authors would like to thank Prof. Joy Paulose, HOD, Department of Computer Science, Christ University, Bangalore, India for constantly motivating us in the entire process of our research.

REFERENCES

Avik Basu, Sanjiban Sekhar RoAvik Basu,Sanjiban Sekhar Roy,Ajith Abraham, 2015. A Novel Diagnostic Approach Based on Support Vector Machine with Linear Kernal For classifying the erythemato-squamous disease. International Conference on Computer Communication Control and Automation.

Bekir Karlik, Gunes Harman, 2013. Computer-Aided Software for Early Diagnosis of Erythemato-squamous Disease. IEEE XXXIII International Scientific Conference Electronic and Nanotechnology (ELNANO).

Beriman, L., 1996. Bagging predictors. Mach Learning.

Chandra, J and Siji T. Mathew,2012. SVM Ensemble Model For Investment Prediction. International Journal of IT, Engineering and Applied Science Research (IJIEASR), pp: 19-23.

Chandra, J., M.Nachimai and Anitha S Pillai, 2015. Predictive Cervical Carcinoma Stages Identification Using SVM Classifier. International Journal of Computer Trends and Technology, pp: 122-125.

Chien-Fu Chien, Shi-Lin Chen and Yih-Shin Lin, 2002. Using Bayesian Network for fault Location on Distribution Feeder. IEEE Transaction on Power Delivery, pp: 785-793.

David Picard, Nicolas Thome and Matthieu Cord, 2013. J Kernel Machines: A Simple Framework for Kernel Machines. Journal of Machine Learning Research, pp: 1417-1421.

Guosheng Wang, 2008. A Survey on Training Algorithms for Support Vector Machine Classifiers. Fourth International Conference on Networked and Advanced Information Management, IEEE.

Haiyi Zhang and Di Li., 2007. Navie Bayes Text Classifier. IEEE International Conference on Granular Computing.

Hatice cataloluk, Metin Kesler, 2012. A Diagnostic Softtware Tool For Skin Disease with Basic and Weighted K-NN, IEEE.

Hemant Palivela, Yogish H.K., S.Vijaykumar and Kalpana Patil, 2013. Survey On Mining Techniques For

Breast Cancer Related Data. Information communication and Embedded System (ICICES), pp: 540-546.

Hoan Tran Quoc, Hideya Ochiai, Hiroshi Esaki, 2015. Hidden Topics Modeling Approach for Review Quality Prediction and Classification. Seventh International Conference of Soft Computing and Pattern Recognition (SoCPaR).

Hongmei Zhang, Quangong Huo and Wei Ding, 2008. The application of AdaBoost-neural network in stored product insect classification. Proceeding of 2008 IEEE international Symposium on IT in Medical and Education.

Imran, K., M. Ture and A.T. kurum, 2008. Comparing performance of logistic regression, classification and regression tree, and neural network for predicting coronary artery disease. Expert system with Applications, pp: 366-374.

Jerome Friedman, Trevor Hasstie and Robert Tibshirani, 2000. Additive Logistic Regression: A Statisical View Of Boosting, pp: 337-407.

Kampouraki, A., D.Vassis, P.Belsis, C.Skourlas., 2013. A Web based Support Vector Machine for Automatic Medical Diagnosis. Proceeding of the 2nd International Conference on Integrate Information, 73:467-474.

Latha Prathiban, R.Subramanian, 2009. An Intelligent Agent for Detection of Erythemato-Squamous Disease using Co-Active Neuro-Fuzzy Inference System and Genetic Algorithm. International Conference on Intelligent Agent & Multi-Agent Systems, IAMA, pp: 1-6.

Li, X., G.C. Nsofor and L.Song, 2009. A comparative analysis of predictive data mining techniques. International Journal of Rapid Manufacture, pp: 50-172.

Moudani, W., 2013. Dynamic Features Selection for Heart Disease. International Science Index, pp: 629-634.

Mythili, T.,Dev Mukherji,Nikita Padalia andAdhiram Naidu, 2013. A Heart Disease PredictionModdelUsingSVM_DecisionTrees-LogisticRegression (SDL). International Journal of ComputerApplication, pp: 0975-8887.

Nicholas, I., Sapankevych and Ravi Sankar, 2009. Time Series Prediction Using Support Vector Machines: ASurvey. IEEE COMPUTATIONALINTELLIGENCE MAGAZINE, 1556-603X.

Omar HEROUANE, Lahcen MOUMOUN and Taoufiq GADI, 2016. Using Bagging and Boosting algorithm for 3D object labeling. International Conference on Information and Communication System (ICICS).

Oztekin, A., D. Delen and Z.J.Kong, 2009. Predicting the graft survival for heart-lung transplantation patients: An integrated data mining methodology. International journal of Medical Informatics (IJMI), 78(12):e84-e96.

Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, 2015. Classification and Prediction Based Data Mining Algorithms to Predict Slow Learning in Education Sector. ICRTC, 57:500-508.

Peng, Y.,Wu Zv andJ.Jiang, 2010. A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 43(1):15-23.

Raymer, M.L.,T.E.Doom, L.A.Kuhn andW.F.Punch, 2003. Knowledge Discovery in Medical and Biological Datasets Using A Hybrid Bayes Classifier/Evolutionary Algorithm. IEEE Transaction On Systems, Man, and Cybernetics, 33: 802-813.

Renders, J.M.and T.Simonart., 2009.Role ofArtificialNeural Networks in DermatologyKARGER, pp: 102-104.

Reyzin,L.and R.Schapire, 2006. How boosting the margin can also boost classifier complexity. in Preceding International Conference, pp: 753-760.

Sasikala, S., A.B.Arockia Christopher, S.Geetha,S. Sppavn alia Balamurugan, 2013. A Predictive Model using Improved Normalized Point Wise Mutual Information (INPMI). Eleventh International Conference on ICT and knowledge Engineering.

Shamsul Arifin, M., M. Golam Kibria, Adnan Firoze, M. Ashraful Amin, Hongyan, 2012. Dermatological Disease Diagnosis Using Color-Skin Images. International Conference on Machine Learning and Cubernetics, Xian.

Shon-Zhong Zhang, Zeng-Nian Zhang, Nan –Hai Yang, Jian-Ying Zhang, Xiu-KunWang, 2004. An Improved EM Algorithm For Bayesian Network Parameter Learning. Proceeding of the Second International Conference on Machine Learning and Cybernetics,Shanghai.

Sho-Zhong Chang Hong yu, H Ua Ding, Nan –Hai Yang-X, Iu-Kun Wang, 2003. An Application of Online Learning Algorithm For Bayesian Network Parameter. Proceeding of the Second International Conference on Machine Learning and Cybernetics.

Souad SOuafi-Bensafi, Marc Parizeau, Franck Lebourgeois, Habert Emptoz, 2002. Bayesian Network Classifiers applied to Documents. IEEE, 1051-4651.

YAYA XIE and XIU LI., CHURN PREDICTION WITH LINEAR DISCRIMENT BOOSTING ALGORITHM. Proceeding of the seventh International Conference on Machine Learning and Cybernetics.

Yuichi Motai., 2015.Association for Classification and Prediction: ASurvey. IEEE.