Decision Tree Classification based Decision Support System for Derma Disease

Garima Sahu
Department of Computer Science & Engg (M.Tech)
Raipur Institute of Technology, (R.I.T.)
Raipur, Chattishgarh, India

Rakesh Kumar Khare
Department of Information Technology (H.O.D.)
Raipur Institute of Technology, (R.I.T.)
Raipur, Chattishgarh, India

ABSTRACT
The process to utilize, the relevant information or knowledge extracted from large databases, into decision making process is called Data Mining. It is widely used in each sector but especially it helps a lot in health care sector so that complicated disease can be diagnosed easily and accurately. In order to diagnose the disease, a decision support system is proposed based upon decision tree technique so that necessary decision can be made after analyzing the input related to the patients. The classification technique which is used to build this model is decision tree, various decision tree based techniques are explored in this study and measured using various measures like accuracy, sensitivity, specificity, precision, recall, F-measure and ROC area. The Dermatology disease is all about the study related to skin disease which is extremely difficult because all six different categories of these diseases share the similar clinical features. The function tree technique is performing very well with overwhelming experiment results of 100 % accuracy, 100% sensitivity and 100 % specificity. The feature selection methods are applied to increase the quickness of the model. With the help of feature selection methods, all the redundant and unwanted features will get removed and a set of effective features will only be required for the purpose of diagnosis of disease. Best first search and rank search are the most suitable feature selection method which can be applied to strengthen the efficiency of the proposed model for derma diseases.

Keywords
Feature selection, Dermatology, Decision tree, Classification.

1. INTRODUCTION
The medical domain requires automated solutions where any possibility of human error exists. In order to remove the dependency upon practical knowledge of the medical practitioner or physicians, the domain requires decision support system which may take necessary action related to diagnosis of diseases and perform the relevant function for the same. It must be capable of prescribing the needed activities to the patients and monitor them. It may issue the alerts, if any discrepancy is found. As dermatology disease is very complex to diagnose because it shares similar features among the different categories of the diseases of same family. An expert system is required for the purpose of diagnosis of disease otherwise the disease may turn out to be in the form of skin cancer.

In order to build the model for the purpose of diagnosis of disease, the classification techniques can be used. Classification [1] is the process to assimilate the data of different categories in one group. This process is very basic one in data mining to put the sample of data of similar type altogether and extraction of information can be performed by applying the set of defined rules. The decision tree classification technique is easily understandable and accurate as well. This technique consists of a set of algorithms which serve the process to classify the data set and select the set of the most relevant features of the dataset. Basically, the model is based upon feature selection algorithm so that diagnosis of disease can be uniform, intelligent and quick [2].

Each algorithm reacts differently towards the training and testing data set so that the split of training and testing data set needs to be considered to build the model and to validate it. These classification techniques evaluate the expert system and produce the results in terms of the system parameters so that usability of the system can be identified clearly. In decision tree classification technique [3], the root node is the top most node of the tree and internal nodes are considered to be the test on attribute. The branch depicts the output of the test performed on attribute and the leaf node represents the class label.

The application of decision tree technique is used to simplify the process of identifying the disease within a short time of span and above all, this model doesn't require previous experience mandatorily. The data is to be split between training set and testing set on the basis of the classifier used. This paper is focused to identify the set of minimum feature which can represent the effect of all features overall. The model responds differently on the application of different classification techniques. To achieve the goal of efficient model for any specific classifier, the split of the health care data set between training and testing set needs to be considered. This model defines the overall efficiency and capability by providing evidences related to its accuracy, sensitivity, specificity, precision, recall, F-measure & ROC area.

Over the years, the expert system for the differential diagnosis of erythemato-squamous has been proposed by incorporating decisions made by nearest neighbor, naive Bayesian and voting feature intervals-5 classifier [4][5].Data mining over medical electronic data from the perspective of characteristics of medical data and requirement of system has been proposed to use the most efficient and effective classification algorithm [6].The expert system for thyroid disease with accuracy of 95.33% has been proposed for the diagnosis[7]. An expert system with the use of neural network, C5.0 decision tree and linear discriminate analysis is suggested for the classification of six different categories of dermatology disease [8]. Neuro fuzzy rules are also one of the medium to use in expert system [9]. The multi attribute decision analysis can also be

performed to apply fuzzy set theory [10]. The enabling factors for the implementation of healthcare information system in the organization can be analyzed well with cross case analysis of pilot trials [11]. To implement the advance features like feature selection from textual sources may increase the scalability of the system [12]. The implementation and development of methods and techniques are well collected at a glance to get the incremental approach in the techniques used in expert system [13].

2. DECISION TREE TECHNIQUE

Decision trees can easily handle the bulk data very easily. These methods are easy to understand and don't require previous acquired knowledge. The process of building the model involves classification by various techniques and application of various rules over that. There is a set of algorithms which are mainly used to form the model as classifier, those algorithms are described below:

- Best-First decision tree used for binary split for both nominal and numeric attributes. in the case of missing values, it used fractional instances.
- Decisionstump performs, the regression like mean squared error and entropy based classification. It can be considered as joint with boosting algorithm. It treats missing value as a new one.
- Functional Tree can be used with vide variety of variables like nominal, binary, numeric and multi class as well at leaves or inner nodes. It is having logistic based regression functions.
- LADTree is based upon logitboost strategy which is used to generate alternate decision tree of multi class type.
- C4.5 accounts for unavailable values, continuous attribute value ranges, pruning of decision trees ad rule derivation. In Building a decision tree, we can deal with training sets that have records with unknown attribute values by evaluating the gain, or the gain ratio, for an attribute by considering only those records where those attribute values by estimating the probability of the various possible results [3].
- LMT algorithm mainly used for building classification trees capable to dealt with nominal, numeric, multi class and missing values at the leaves with the logistic based regression functions.
- RandomForest algorithm is used to create forest, with selected number of features, selected via rank based algorithm or principal component analysis, of random trees.
- RandomTree doesn't performs pruning and create trees with N attributes which are chosen randomly at each node. It may allow the estimated probability of class depending upon holdout set.
- REPTree deals with missing values similar to C4.5 and performs pruning of tree on basis of reduced error. The sorting of numeric attributes can be done only once.
- NBTree algorithm generates a decision tree which consists of naive bayes classifiers at leaf node.

3. HEALTH CARE DATA

The health care data which is used in this model has been taken from UCI machine learning dataset. UCI[14] is a machine learning repository which consist the valid data sets related to dermatology. The dataset related to derma disease is mainly classified according to six categories of diseases of dermatology family. This data set consists of overall thirty five features; in which last feature is class or target and rest of the features are known as input for the model. The dataset consists of 365 instances out of which the distribution is as follows: 31% instances of psoriasis, 16% instances of class seboreic, 19% of lichen class, 13% instances are of pityriasis class, 14% are of class chronic and rest instances are of class named as rubra pilaris. The data set covers all the features which are diverse in nature so that model can be tested with the any type of instances and the results should be accurate for the purpose of monitoring the patients. The classification techniques can be applied on the basis of target feature. The UCI repository is the reliable source of the health care data so that the results coming out of the experiment may reflect the scalability of the model.

4. EXPRIMENTAL FRAMEWORK FOR HEALTH CARE DATA CLASSIFICATION

Decision Support System [15] is a tool which helps humans to increase the efficiency of their decision making capability in the field of different complex domains such as health.

In this paper, the emphasis is provided upon clinical decision support system which is a computer software developed to assist staff, physicians and patients as well in their routine clinical decisions. It acts like a bridge between the knowledge of health to influential health options for the purpose to provide improved health care to the patients.

The decision support system (see Figure 1) is an integrated system which takes the input and analyzes it by applying the set of rules and provides accurate and efficient results quickly.

4.1 Components of DSS

4.1.1 Knowledge Source (Data Set)

It is the collection of structure less expertise of solving the complex problems. It consists of real time data related to any specific problem and use it to identify the condition specific solution of the problem. It helps to gather the information related to the patient. The different knowledge sources can be the reports, books, case studies, practical experience of medical practitioner or the relevant patient data suffering from the specific disease.

Knowledge source is divided into Training data set and testing data set. The training data set is used to build the model and testing data is used as input for knowledge base for the purpose of validation of results produced by the model.

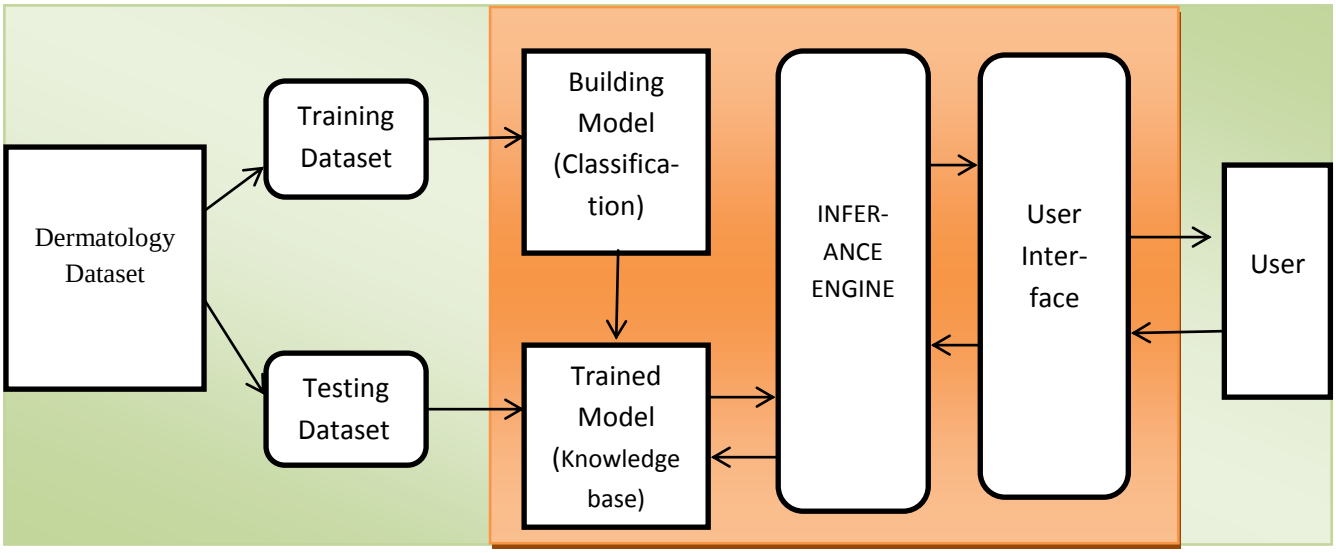

Fig 1: Architecture of Decision Support System

4.1.2 The Trained Model (Knowledge Base)

The trained model consists of all the properties which are validated while building the model, with the training data. Knowledge base acts like a brain for processing the testing data with utmost accuracy and efficiency. Some of the rules and facts won't change themselves while compilation and execution of the data. There are some facts which are about the specific consultation of the complete system. In the operational phase, these facts expand themselves to produce different decisions along with the static knowledge. Overall knowledge base is having related knowledge which helps a lot to solve and understand the problems.

4.1.3 Inference Engine

It involves a mechanism to provide reasoning about information present in the knowledge base and formulate the conclusions using control strategy.

Inference means search through knowledge base and extract new knowledge. In order to provide formal reasoning and unification, this process involves matching order sets. This function is similar to the human experts which are used to solve problems in various knowledge domains.

Inference operates by using rules. Its control strategy performs the ordering of rules which are required to be applied. Backward chaining and forward chaining are the two types of control mechanism in the systems.

4.1.4 User Interface:

The system consists of a language processor for problem, condition-specific or friendly communication between user and the computer. The user interface is used to form the communication channel in a natural language.

The user is integral part of DSS, which use the system and get benefited out of it. User interface is the part of the system with which a user may interact.

All these components perform their tasks to provide the decisions to the users. The cohesiveness of all the components determines the quality and efficiency of the decisions produced by the system.

5. EXPERIMENTAL SETUP

The experiment work is carried out using WEKA open source data mining toll under windows environment and start with the training set and testing set. The whole set consists of numerous features related to the disease. The data related to the features had been collected by the patients suffered with the disease and the data is available at UCI repository. There are overall thirty five features which are available for the purpose of experiment, in order to increase the productivity and diversity of the experience.

The application of different classifiers is being performed to evaluate the model on the basis of accuracy, sensitivity and specificity etc. The experiment results for respective decision tree classifying technique are presented (see Table 1). Many decision tree techniques presented (see Table 1) are inbuilt in WEKA, user can use these techniques by setting its parameters to develop model to diagnose dermatology related disease. It is evident that the model is providing efficient results so that it adds great help in the process of diagnosis of the dermatology disease.

Table -1 Experiment results after application of decision tree classifying techniques

S. No	Model	Accuracy		Sensitivity		Specificity		Precision		Recall		F-Measure		ROC area	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
1	Bftree	98.3607	95.9677	98.4	98.4	99.7	99.7	98.4	96.1	98.4	96	98.4	96	99.7	98.3
2	DecisionStump	51.2097	52.9915	51.2	53	86.3	86.2	35.8	32.7	50.6	53	39.3	39.2	76.1	80.7
3	FT	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
4	J48	97.8903	92.9688	97.9	93	99.5	97.5	98.4	93.1	98.4	93	98.4	92.3	99.7	96.4
5	J48graft	97.8903	92.7419	97.9	92.7	99.5	96.9	98.4	93.4	98.4	92.7	98.4	92.6	99.7	96.9
6	LADTree	99.1837	98.3471	99.2	98.3	99.8	99.6	99.2	98.4	99.2	98.3	99.2	98.3	100	99.9
7	LMT	100	98.2906	100	98.3	100	99.7	100	100	100	100	100	100	100	100
8	RandomForest	100	98.3871	100	98.4	100	99.7	100	98.4	100	98.4	100	98.4	100	99.6
9	RandomTree	100	93.6364	100	93.6	100	99	100	94.3	100	93.6	100	93.8	100	96.3
10	REPTree	92.6641	89.6226	92.7	89.6	97.7	97.6	95.2	89.4	95.1	89.6	95.1	89.2	99	96.7
11	SimpleCast	96.4844	95.4545	96.5	95.5	99	98.8	96.4	95.6	96.3	95.5	96.3	95.5	99	98
12	NBTree	98.4127	100	98.4	100	99.6	100	98.4	100	98.4	100	98.4	100	100	100

Accuracy, sensitivity and specificity at training and testing stages are also shown in bar graph respectively (see Figure 2) (see Figure 3) (see Figure 4). It is clear from these figures that functional tree is outperforming with 100% efficiency as compare to other decision tree techniques. The models depict differently with the testing and training data, it proves the diversity of the instances collected in the dermatology data set which is used for the experiment.

Table -2 Effect of feature selection method

Sr. No	Model	Feature Selection Method	Total no. of feature available in dataset	Total No. of effective Feature	Total no. Of Reduce feature
1	Function Tree	Best First Search	34	21	13
2	Logistics Model Tree	Rank Search	34	17	17

On applying the feature selection method over the classification technique based model, it helps to find out the effective features available in the data set (see Table 2). The effective features are those which are essentially required to be analyzed to get the result of the diagnosis in the data set, there can be features which are generating same effect so only one feature is sufficient to represent all the other features. Similarly few of the inactive features also exist which doesn't affect the diagnosis process. On applying the best first search method, it helps to sustain the accuracy, specificity & sensitivity of the model even after reducing thirteen features. The efficiency of the model is definitely increased through this. Several feature selection methods are available but each model doesn't support positively towards every method. The Rank search feature selection method is very useful in case of Logistic Model Tree based model because the accurate diagnosis can be performed with only seventeen features, even though the original data set of patient data consists of overall thirty four features. This depicts the importance of usage of the feature selection methods with the decision tree classification technique based decision support system.

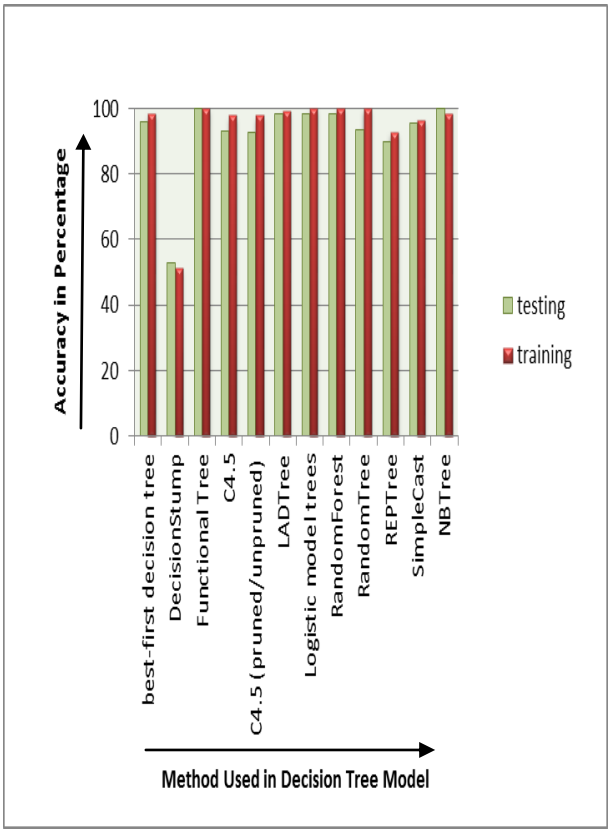

Fig 2: Accuracy of models with testing & training data set

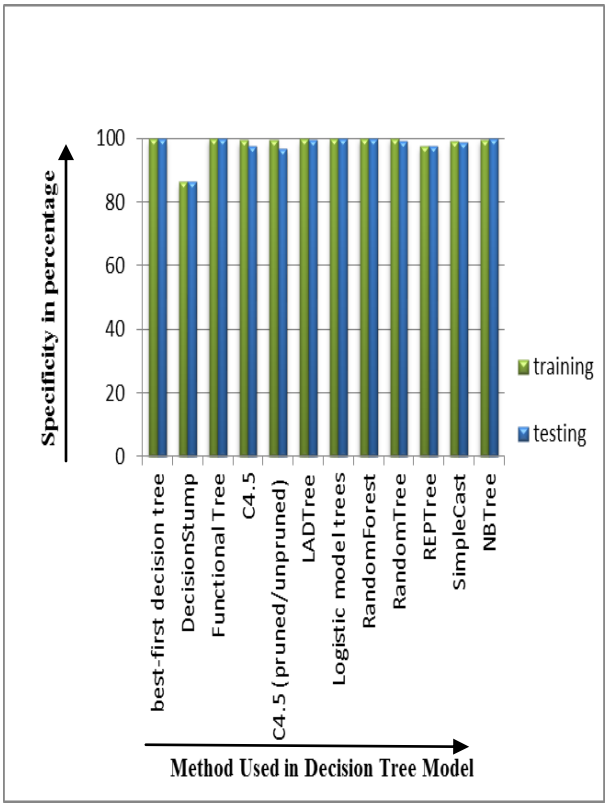

Fig 3: Specificity of models with testing & training data set

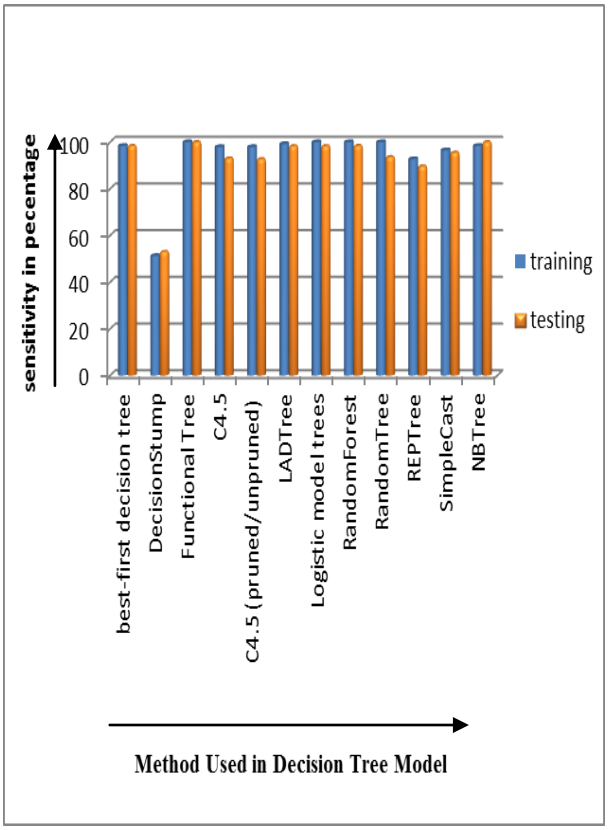

Fig 4: Sensitivity of models with testing & training data set

6. CONCLUSION

The experiments and its results determine the great scope in the direction of latest advancement in diagnosis of disease. This paper mainly deals with Decision tree based classification techniques and their effect upon model to evaluate it. There are many other classification techniques for which a different and more diverse health care data set may give amazing results which will definitely help in the research of modern health care sector. Through this paper, set of minimum features that may be sufficient to represent the whole set is tried to be identified. Further development suggests towards the new algorithm which may be useful to select the feature and help to form the set of features which may represent the effect of all the available features accurately and efficiently. This model, along with other efficient models based upon different classifying techniques, can be incorporated to prepare a web based integrated solution for the diagnosis of any disease.

7. REFERENCES

[1] Sumathis and Paneerselvam, surekha (2010), computational Intelligence Paradigms, Theory and application using MATLAB, CRC Press , 52-54 Boca Raton , PL

[2] Han, J., Kamber, M., and Pei, J. (2011).Data mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, San Francisco, CA. USA.

[3] Pujari,A,K , (2012), Data Mining Techniques, 2nd Edition , Universities Press (India) Private Limited, Himayatnager, Hyderabad 500029 (A.P.)

[4] H.A. Guvenira, N. Emeksizb,(2000), An expert system for the differential diagnosis of erythemato-squamous diseases, Science-Expert systems with applications 18, 43-49

[5] Guvenir.H.Altay,Demiroz.Gulsen,Ilter.Nilsel,(1998),Lea rning differential diagnosis of erythemato-squamous disease using voting feature intervals,Elsevier,Artificial Intelligence in Medicine,13,147-165.

[6] AI-Aidaroos.k.m, A.A.Bakar, Z.Othman (2012), Medical data classification with Navive Bayes Approch,Asian network for scientific information, Information Technology journal, 11(9), 1166-1174.

[7] Keles.Ali, Keles.Ayturk & Yavuz.Ugur,(2011),Expert system based on neuro-fuzzy rules for diagnosis breast cancer,Elsevier, Expert system with applications,38,5719-5726

[8] M.Elsayad.Alaa, (2010), Diagnosis of Erythemato-Squamous diseases using ensemble of data mining methods,ICGST-BIME journal,Volume 10,Issue 1

[9] Keles.Ali,Keles.Ayturk,(2008),ESTDD:Expert system for thyroid diseases diagnosis,Elsevier,Expert systems with applications,34,242-246

[10] Yan.Hong-Bin,Huynh.Van-Nam, Ma. Tieju, Nakamori,Yoshiteru, (2013), Non –additive multi –attribute fuzzy target-oriented decision analysis, Elsevier , Information science ,240, 21-44.

[11] Yang.Zienbin, Kankanhalli, Ng.Boon-Yuen, Yong.Lim. Tuang.Justin, (2013), Analyzing the enabling factor for the organizational decision t adopt healthcare information system, Elsvier, Decision support system, 55, 764-776.

[12] Vicient.Carlos,Sanchez.David,Moreno.Antonio,An automatic approach for ontology-based feature extraction from heterogeneous textual sources,Elsevier,Engineering applications of artificial intelligence

[13] Liao.Shu-Hsien,(2005),Expert system methodologies and applications-a decade review from 1995-2004, Elsevier, Expert system with applications, 28, 93-103

[14] UCI (2012). Web source: http://archive.ics.uci.edu/ml/datasets.html,last accessed on Jan 2012.

[15] H.Dunhan, Margaret (2009), Data mining –Introductory and Advance Topics, Sixth Edition , Dorling Kindersley (India) Pvt. Ltd, New Delhi, India