Survey on Classification Methods using WEKA

Meenakshi
Software Engg
ITM University
Gurgaon

Geetika
Computer Science and Engg
ITM University
Gurgaon

ABSTRACT
In data mining classification is to accurately predict the target class for each case in the data. Decision tree algorithm is one of the commonly used classification algorithm to make induction learning based on examples. In this paper we present the comparison of different classification techniques using WEKA. The aim of this paper is to investigate the performance of different classification methods on clinical data. The algorithm tested are Bayes Network, Navie bayes, Logistic, rule jrip, and J48.

Keywords
Data classification, Information gain, Decision tree, Weka.

1. INTRODUCTION
Data mining can be used in context of many things such as pattern recognition, statistics, and database management and computer science. It can be used to discover relation between large sets of data. Data mining can be expressed as it is a process of indentifying easily understandable patterns with the help of identifying novels and point, thus providing the meaningful information from the provided data set. While using it in database it involves the prediction of variables or identifying fields and using them in a better and meaningful manner for predicting future values and behavior. Now a days as healthcare system is a vast area for researchers for work on so data mining is used in it to improve the quality of health care by reducing costs. Health care systems dataset includes physician practices, disease management and resource utilization.

CLASSIFICATION USING WEKA
Weka is written in java and can run on any of the platform. We can say that Weka is a collection on of algorithms with the help of which real world problems can be solved. Algorithms can be applied either directly or to a dataset called from own java code. Data processing, classification, clustering, visualization regression and feature selection these techniques are supported by Weka. In Weka data is considered as an instances and features as attributes [6]. In this main user interface is the explorer but essential functionality can be attained by component based knowledge flow interface and command line whenever simulation is done than the result is divided into several sub items for easy analysis and evolution. One part in correctly or correctly classified instances partitioned into percentage value and numeric value and subsequently kappa statistics mean absolute error and root mean squared error will in numeric value.

In data mining, an important problem is large data set of classification. For a database with a set of classes and number of records such that each record belongs to one of the given classes, the problem of classification is to decide the class to which a given record belongs. Here, it is concerned with a type of classification called supervised classification. In supervised classification, a training data set of records and for each of this set, the respective class to which it belongs is also known. As they represent rule, Decision tress are especially attractive in data mining environment.

1.1 C4.5
C4.5 developed by rossouinlan and is an extension of ID3 it is also known as statistical classifier. With the help of training dataset decision tree is being constructed and entropy is being computed or calculated up. Training dataset is given as S= {S_1, S2, S3……} are classified samples. S_1 consists of P dimensional vector (X_1i, X2i, …, .X_pi).here Xi is attribute and a class n which Ni falls. As we constructs a decision tree in C4.5 and every time node of the tree is chosen and is subdivided in to subsets. Normalized information gain is the condition on which division is based. Attribute having the highest normalized gain is chosen to take decision tree. C4.5 helps in reducing error, chosen suitable attributes etc.

1.2 ID3
It is extended version of C4.5, Quinlan introduced the ID3, Iterative Dichotomizer 3, for constructing the decision trees from data. To measure how informative is a node, entropy is used. ID3 algorithm uses the criteria of information gain to determine the goodness of a split [2]. The attribute with the greatest gain is taken as the splitting attribute, and the data set is split for all distinct values of the attribute. Shortcomings of ID3:
1) It often biased to select attributes with more taken values, which are not necessarily the best attributes.

2) ID3 algorithm selects attributes in terms of entropy which is computed based on probabilities, while probability method is only suitable for solving stochastic problems.

1.3 SVM
Algorithm is used to solve the optimization problem is known as sequential minimal optimization. Works by breaking optimization problem in to several sub problems and then solved analytically multiplier a1which is larger has linear quality constraints and thus the smallest problem has two multipliers. For multiplier constraints produced are $0<a_1$, $a2<c$, $y1a1+y_2a_2$.

1.4 Navie Bayes
When the dimensionality of the inputs is high, the Naïve Bayes Classifier technique is particularly suited. The problem with the Naïve Bayes Classifier is when it assumes all attributes are independent of each other which in general cannot be applied. Naive bayes is harder to debug and understandable [2]. Naive bayes used in robotics and computer vision. In naive bayes decision tree perform poorly.

Naive Bayes is a probabilistic based classifier which applies Bayes' theorem (or Bayes's rule) with strong independence (naive) assumptions.

P (H/E) = P (E/H)*P (H)
Bayes's rule determines that the outcome of a hypothesis or an event can be predicted based on observations of some evidences. From Bayes's rule, we have

(1) A priori probability of H or P (H): This is the probability that an event occurred before the evidence are observed.

(2) A posterior probability of H or P (H/E): This is the probability that an event occurred after the evidence are observed.

Naive Bayes has an advantage that it requires small training data, while estimating parameters (means and variances of the variable) which are necessary for classification, this is because independent variables are assumed to be the variances of variables as each class needs to be determined and not the complete covariance matrix.

1.5 Bayesnet

Bayesian network is used for classification and received a considerable attention. It is a powerful probabilistic representation. Classifier learns from the training data. Here Ai is the conditional probability of each attribute and a class label C. In this classification is done by computing the probability of c for particular A1….An and predicting the class with highest posterior probability. Goal is to predict class a vector of predictors or attributes.

Bayesnet are a powerful reasoning mechanism and knowledge representation. Bayesnet represent causal and events relationships between them as conditional probabilities involving random variables. Given the values of a subset of these variables (evidence variables) Bayesnet can compute the probabilities of another subset of variables (query variables). Bayesnet can be created automatically (learnt) by using statistical data (examples). The well-known Machine Learning algorithm, Naïve Bayes is actually a special case of a Bayesian Network.

1.6 Logistic regression

Logistic Regression is a predictive model which is used when the target variable can also be used as categorical variable [7]. The variables have two categories like live/die, have disease/doesn't have disease, purchases product/doesn't purchase, wins race/doesn't win, etc. Decision trees are not present in a logistic regression model, whereas logistic regression model is more into nonlinear regression e.g. putting polynomial to set of data values.

Logistic regression uses two types of target variables:

(1)Categorical target variable consist of two categories of variables that are binary and dichotomous variable.

(2)Continuous target variable consist of values that ranges from 0.0 to 1.0, which are used to represent probability values or proportions.

The logistic formula consists of continuous predictor variable. Dichotomous predictor variable has values of 0 or 1, and with a dummy variable for each and every category of predictor variables.

The logistic model formula is:

$P=1/ (1+\exp (-(B_{0+}B_{1*}X_{1+}B_{2*}X_{2+\ldots.}B_{k*}X_{k})))$

B_0 refers to a constant and B_i represents coefficients of predictor variables (or dummy variables as in case of multi-category predictor variables). The value computed, P, is a probability which ranges from 0 to 1. B_0 constant can be excluded by turning off the option "Include constant term" which is present on the logistic regression model property page.

For predicting the values of the dependent variable, logistic regression estimates the probability.

For eg: predict whether the patient diseases or not. With the help of logistic regression estimates that probability of patient having disease. If the estimated probability is greater than 0.5 than there is a high probability of patient having the diseases function is not a linear function can be represented as?

$P(y) = 1/\{1 + e(-a - bx)\}$

P(y) =Probability that y, y=dependent variable occurs, x=value of the attribute, a=constant, b=coefficient of the independent variable.

1.7 Jrip rule classifiers

Jrip (RIPPER) is one of the most popular algorithms; it has classes that are examined in increasing size. It also includes set of rules for class is generated using reduced error Jrip (RIPPER). Proceed by treating examples of judgments made in training data as a class, and finding rules that covers all the members of the class. Then it proceeds to the next class and repeats the same action, repetition is done until all classes have been covered.

1.8 J48

It is the classifier according to which we classify our classes it is also known as free classifier who accepts nominal classes only. In this prior knowledge should be there while classifying instances. It is used in the construction of decision tree from a set of labeled training data using the information entropy. Attributes which we use helps in building decision tree by splitting it into subset and normalization information gained can be calculated. Splitting process comes to an end when all instances in a subset belong to the same class. Leaf node is also is also present or being created to choose that class a possibility also can be there that none of the feature provides information gain.J48 creates decision nodes up higher in the tree using expected value of the class.J48 can use both discrete and continuous attributes, attributes with differencing lost and training data with missing attribute values.

2. CONFUSION MATRIX

Metrics are measures that we can analysis and control without prior assumption about the data metrics relates a classifier to the process that produce the data without relying on the data itself [3].

Fig 1. Table of Confusion Matrix

	Predicted-ve	Predicted +ve	Total
Actual -ve	True +ve	False +ve	N
Actual +ve	False -ve	True +ve	P
Total	n	p	

Overall error=FP+FN/P+N$\frac{FP+FN}{P+N}$

Accuracy =TP+TN/P+N

3. DATASET FOR CLASSIFICATION

All the classification algorithm are applied on clinical data set as given in table 1[4]. Weka tool has been used for the analysis.

Table 1. Dataset of Healthcare sample

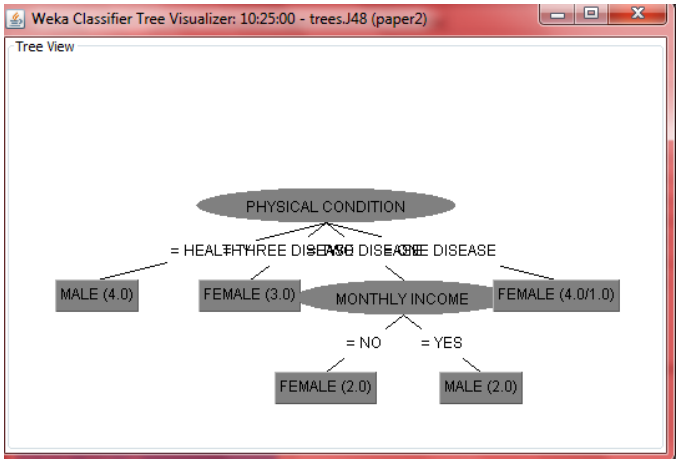
Property	Monthly income	Physical condition	Gender	Classification
One	No	Healthy	Male	N
Two	No	Suffering three disease	Female	P
Three	No	Healthy	Male	N
Four	Yes	Suffering two disease	Male	P
Five	No	Suffering three disease	Female	P
Six	No	Suffering one disease	Female	N
Seven	Yes	Suffering two disease	Male	P
Eight	No	Suffering one disease	Male	N
Nine	No	Suffering two disease	Female	N
Ten	No	Suffering two disease	Female	P
Eleven	No	Healthy	Male	P
Twelve	No	Suffering one disease	Female	P
Thirtheen	No	Suffering three disease	Female	P
Fourteen	No	Healthy	Male	N
Fifteen	Yes	Suffering one disease	Female	P

Table 2. Information Gain of Dataset shown in table 1

Attribute	Information gain
Monthly income	0.17
Physical income	0.18
Gender	0.09

Decision tree

Fig 2. Decision tree for Healthcare Data shown in Table1



Decision tree depends upon information gain. As per table of information gain physical condition have maximum value so that division will occurs initially on physical condition then monthly income and so on.

RESULTS

Table 3. Simulation result of each algorithm

Algorithm Total instances (15)	Correctly instances% (value)	Incorrectly instances%(value)	time taken	Kappa statics
Bayes net	93.333% (14)	6.6667% (1)	0	0.8649
Navie bayes	86.667% (13)	13.333% (2)	0.02	0.7222
Logistic	80% (12)	20% (3)	0.11	0.5714
Rule jrip	73.333% (11)	26.667% (4)	0	0.4118
J48	60% (9)	40% (6)	0	0

Table 4. Training and Simulation error

Algorithms (Total instances 15)	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)
Bayes net	0.153	0.2415	31.7224	49.2793
Navie bayes	0.2127	0.2808	44.0874	57.2963
Logistic	0.2333	0.3416	48.374	69.7016
Rule jrip	0.3909	0.4421	81.0421	90.2178
J48	0.48	0.4899	99.5122	99.9712

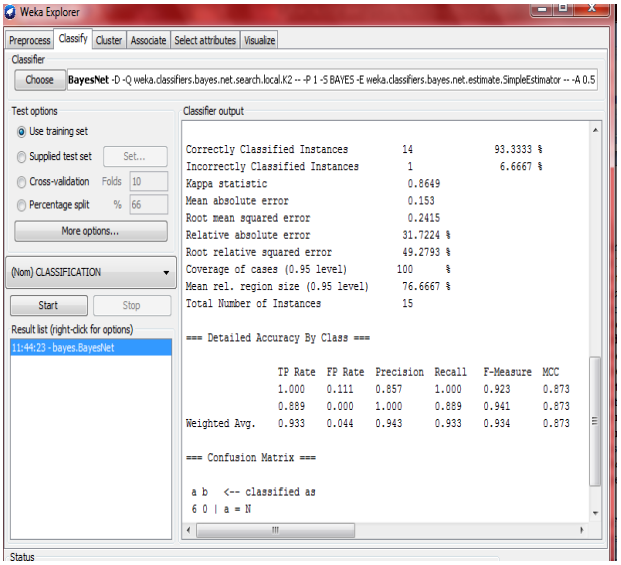


Fig 3. Table of Classify Bayesnet in Weka

The correctly/incorrectly classified instances define the case where the instances are used as test data. Kappa statistics is a measure of agreement normalized for chance agreement.

K=P (A)-P (E)/1-P (E).

Where P (A) is the percentage agreement (e.g., between your classifier and ground truth) and P (E) is the chance agreement. K=1 indicates perfect agreement, K=0 indicates chance agreement.

The mean absolute error is the sum over all the instances and their Abs Err per Instance divided by the number of instances in the test set with an actual class label.

Mean Abs Err = Sum (Abs Err Per Instance)

Root Relative Squared Error is computed by dividing the Root Mean Squared Error by the Root Mean Square error obtained by just predicting the mean of target values (and then multiplying by 100). Therefore, smaller values are better and values > 100% indicate scheme is doing worse than just predicting the mean. Root Absolute Error is computed in similar manner.

4. CONCLUSION

The accuracy of Bayes net is highest as 93.33% they are fast and easy to implement as they depend on class conditional independence. Bayes network classifier has the lowest average error as compared to others. Decision tree are easy to understand and the tree generated can be easily transformed to if then else rules but finding an optimal solution using decision tree is NP problem though appropriate attribute selection in decision tree can improve its performance.

5. REFERENCES

[1] Geetika "A Survey of Classification Methods and its Application "2013.

[2] Khalid Raza and Atif N. Hasan. "A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microarray Data." arXiv preprint arXiv:1307.7050 (2013)..

[3] Mr V.K pachghare,Parag Kulkarni,"Pattern based network security using Decision Tree and Support vector Machine"IEEE,2011.

[4] Remco R. Bouckaert, Eibe Frank "WEKA Manual For Version 3-7-4 University of Waikato" Hamilton, New Zealand, June 2011.

[5] Linna li, Xuemin Zhang"Study of data mining algorithm based on decision tree"Iinternational conference on computer design and application: 2010.

[6] Mohd Fauzi bin Othman, Thomas Moh Shan Yau. "Comparison of different techniques" University technology Malaysia, skudai, Malaysia.2010.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes "The WEKA Data Mining Software: An Update**"** University of Waikato Orlando, New Zealand.2009.

[8] Wen Zhanga, Taketoshi Yoshidaa,Xijin Tangb"Text classification based on multi-word with support vector machine Knowledge-Based Systems" Elsevier Volume 21, Issue 8, December 2008.

[9] J. Michael Hardin and David C. Chhieng. "Data Mining and Clinical Decision Support Systems".2007.

[10] Shahabi C, Zarkesh A M, Adibi J, et al " Introduction of neutral network [C] "B inningham: IEEE Press,2001.

[11] Mobasher B, Cooley R, Jaideep S, etal "Comments on decision tree" New York: IEEE Press, 1999.

[12] Nils J." Introduction to Machine Learning" California. United Stated of Americas. Nilsson (1999).

[13] C. X. Ling and C. Li "Data mining for direct marketing Specific problems and solutions" In Proceedings of FourthInternational Conference on Knowledge Discovery and Data Mining (KDD-98), pages 73 – 79. 1998.