Volume 4, No. 4, April 2013

Journal of Global Research in Computer Science

ISSN-2229-371X

REVIEW ARTICLE

Available Online at www.jgrcs.info

EVALUATION OF CLASSIFICATION ALGORITHMS FOR DISEASE DIAGNOSIS

Tamije Selvy P[1], Palanisamy V[2], Elakkiya S[3*]
[1]CSE, Sri Krishna Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Principal, Info Institute of Engineering, Coimbatore, Tamil Nadu, India
[3]CSE, Sri Krishna Institute of Technology, Coimbatore, Tamil Nadu, India
elakkiya.soundar@gmail.com[3]

Abstract: Data classification is the categorization of data for its most effective and efficient use. Data can be classified according to any criteria, not only relative importance or frequency of use. Classification plays a major role in disease diagnosis. The paper contains brief discussion of various classification methods that includes Case Based Reasoning, decision trees, K-nearest neighbour classifier, naïve bayes classifier and neural network. The paper also discusses some applications of classification model. The performance of the classification methods are observed where the CBR classification model results in 90.7% of specificity, 92.3% of sensitivity and 95.5% of prediction accuracy.

Keywords: Classification, disease diagnosis, Case Based Reasoning, decision trees, K-nearest neighbour classifier, naïve bayes classifier, neural network, accuracy

INTRODUCTION

Data Mining refers to the extraction of previously unknown and potentially useful information from data in databases. Data Mining is a part of knowledge discovery process. It is a clever technique that can be applied to extract useful patterns. In addition to collecting and managing of data, data mining also includes analysis and prediction.

Classification techniques in data mining are capable of processing a large amount of data. It can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data. Thus it can be outlined as an inevitable part of data mining and is gaining more popularity [2].

In the present paper a study on various classification techniques have been made. Next section deals with a study on k-nearest neighbor mechanism. decision tree, deals with Bayesian network, deals with Case Based Reasoning and describes neural networks. Finally last section concludes the paper.

K-NEAREST NEIGHBOUR

Nearest neighbor (NN) [1] also known as Closest Point Search is a mechanism that is used to identify the unknown data point based on the nearest neighbor whose value is already known. It has got a wide variety of applications in various fields such as Pattern recognition, Image databases, Internet marketing, Cluster analysis etc.

Nearest Neighbor mechanism can be classified into two types. They are Structure based and Structure less NN classification techniques. K-NN comes under the structure less classification technique . Structure based deals with the basic structure of the data where as structure less mechanism is associated with training data samples. Latter overcomes the memory limitation whereas the former reduces the computational complexity. It makes use of the more than one nearest neighbor to determine the class in which the given data point belongs to and hence it is called as K-NN.

These data samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique. All these data points are necessary in order to make a decision in determining the class of the given data point. There are a large number of machine learning algorithms and K-NN is the most simplest among them. It can also be considered as the one among the top ten data mining algorithms.

K-NN basically works on the assumption that the data is contained in a feature space. Hence all the points are contained in it, in order to find out the distance among the points Euclidian distance or Hamming distance is used according to the data type of data classes used. Here a single number k is given which is used to determine the total number of neighbors that determines the classification. If the value of k=1, then it is simply called as nearest neighbour [3]. K-NN requires:

a. An integer k
b. A training data set
c. A metric to measure closeness

Following fig 1 shows how the classification can be done based on the value of k.

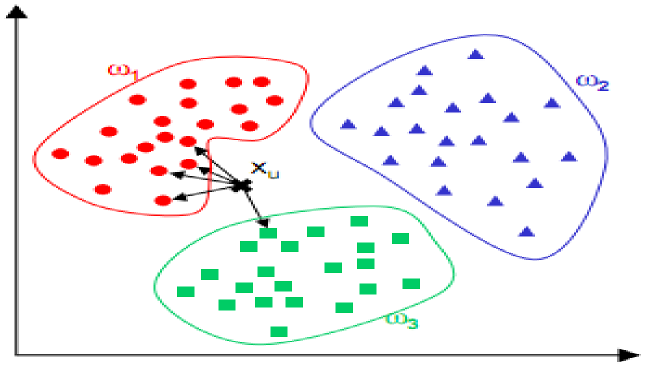


Figure 1. Example of K-NN classification

K-NN mechanism is easy to implement and hence it makes the implementation and debugging process to be faster. It can also help in easy analysis of the neighbour points. Hence the major advantage of this method is that training can be done in a faster manner, simple and easy to learn. Large training data can be determined and hence is a robust mechanism. It basically focuses on large training data sets. Several noise reduction techniques can be used that can be used to improve the classifier mechanism.

Some of its disadvantages are its memory dependency, computational complexity and also its reliance on k-value. It also requires large computational time and hence is a slow technique since all process is done during the run-time.

DECISION TREE

Decision Tree (DT) is another classification technique which is commonly used. It is constructed by examining a set of training samples whose class labels are known. Then these features of known samples are applied in order to determine the properties of unknown samples. They can be regarded as a powerful and popular tool for classification and prediction process. Key requirements for constructing a decision tree are its attribute-value description which means its objects should be expressible in terms of a fixed collection of points called attributes, predefined classes also called as the target classes which have discrete output values [4] and finally sufficient data which helps in understanding the model completely.

Decision Tree is a classifier which has the form similar to that of a tree and has the following structure elements:

a. Root node: Left-most node in a decision tree
b. Decision node: Specifies a test on a single attribute
c. Leaf node: Indicates the value of target attribute
d. Edge: Split of an attribute
e. End-point: Right most node representing final outcome

DT is constructed using divide and conquer (D&C) approach. Each path in DT determines a decision rule. Usually it follows a greedy approach from top to bottom ie; from root node to the ending node recursively for determining the final outcome and hence can deal with uncertainties [5]. D&C strategy approaches a problem in the following manner:

a. Breaking the problem into different sub-problems which are the instances of the given problem
b. Recursively solving each of these problems
c. Finally combining each answers of these sub-problems into a single one

Decision Tree can be considered as more interpretable compared to that of neural networks and support vector machines since they combines more data in an easily understandable format. Even small changes in the input data may lead to great variations in constructing the DT. In some cases it has to deal with uncertainties. This can be solved using sequential decision making of DT. The process of determining the expected values from the end node back to the root node is known as decision tree roll-back.

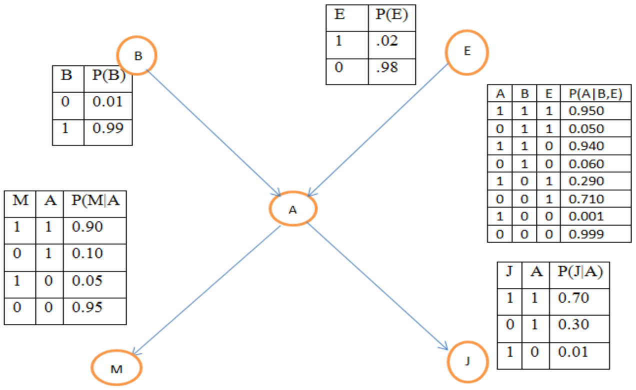
Decision Tree can be explained with an example as given below. Usually DT follows a top-down approach. In the example it shows a weather forecasting methodology which deals with predicting whether it is sunny or rainy and what about the humidity if it is sunny. Thus this can be applied to determine whether the climate suits well for playing tennis. Hence one can easily determine the present climate as well as what will be followed by in the future and based on that the decision can be made if the match can be held or not. This can also be applied in various other applications such as rolling a die, product decision, etc for prediction analysis.

Some of the advantages of DT are they are computationally cheap, easy to use and implement and simple. It also provides objective analysis to decision making, allows flexibility and effective for decision making. Major drawback of DT is that the whole process relies on the accuracy of the input data used and also requires qualitative data to determine the accuracy of the output.

BAYESIAN NETWORK

It provides new ways of exploring and understanding data. It learns from the “evidence” by calculating the correlation between the target i.e. dependent and other i.e., independent variables. The probabilistic model of NBC is to find the probability of a certain class given multiple disjoint (assumed) events. The Naïve Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f(x) can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values <a1, a2, an>. [11]The learner is asked to predict the target value, or classification, for this new instance. A Bayesian network is a representation of the joint distribution over all the variables represented by nodes in the graph. Let the variables be X(1), ..., X(n). Let parents (A) be the parents of the node A, then the joint distribution for X(1) through X(n) is represented as the product of the probability distributions P(Xi|Parents(Xi)) for i = 1 to n. If X has no parents, its probability distribution is said to be unconditional, otherwise it is conditional. The conditional probability values of all the attributes with respect to the class are pre-computed and stored on disk. This prevents the classifier from computing the conditional probabilities every time it runs.

This stored data can be reused to reduce the latency of the classifier [8]. The most interesting feature of Bayesian Networks, compared to decision trees is most certainly the possibility of taking into account prior information about a given problem. An important advantage of Naive Bayes is that the simple structure lends itself to comprehensible visualizations. Bayesian networks can readily handle incomplete data sets. Bayesian networks allow one to learn about causal relationships Bayesian networks readily facilitate use of [6] prior knowledge. Bayesian classifiers are used when the data is high, the attributes are independent of each other and when we want more efficient output, as compared to other methods output. Fig 2 shows the Bayesian Network.



E	P(E)
1	.02
0	.98

B	P(B)
0	0.01
1	0.99

A	B	E	P(A|B,E)
1	1	1	0.950
0	1	1	0.050
1	1	0	0.940
0	1	0	0.060
1	0	1	0.290
0	0	1	0.710
1	0	0	0.001
0	0	0	0.999

M	A	P(M|A
1	1	0.90
0	1	0.10
1	0	0.05
0	0	0.95

J	A	P(J|A)
1	1	0.70
0	1	0.30
1	0	0.01

Figure 2. Bayesian Network

CASE BASED REASONING

Case-based reasoning can mean adapting old solutions to meet new demands; using old cases to explain new situations; using old cases to critique new solutions; or reasoning from precedents to interpret a new situation (much like lawyers do) or create an equitable solution to a new problem (much like labor mediators do). If we watch the way people around us solve problems, we are likely to observe case-based reasoning in use all around us. Attorneys are taught to use cases as precedents for constructing and justifying arguments in new cases [7]. Mediators and arbitrators are taught to do the same. Other professionals are not taught to use case-based reasoning, but often find that it provides a way to solve problems efficiently. Consider, for example, a doctor faced with a patient who has an unusual combination of symptoms. If he's seen a patient with similar symptoms previously, he is likely to remember the old case and propose the old diagnosis as a solution to his new problem. If proposing those disorders wastime-consuming previously, this is a big savings of time. Of course, the doctor can't assume the old answer is correct. He/she must still validate it for the new case in a way that doesn't prohibit considering other likely diagnoses. Nevertheless, remembering the old case allows him to generate a plausible answer easily. Similarly, a car mechanic faced with an unusual mechanical problem is likely to remember other similar problems and to consider whether their solutions explain the new one. Doctors evaluating the appropriateness of a therapeutic procedure or judging which of several are appropriate are also likely to remember instances using each procedure and to make their judgements based on previous experiences. Problem instances of using a procedure are particularly helpful here; they tell the doctor what could go wrong, and when an explanation is available explaining why the old problem occurred, they focus the doctor in finding out the information he needs to make sure the problem won't show up again. We hear cases being cited time and again by our political leaders in explaining why some action was taken or should be taken [10]. And many management decisions are made based on previous experience. Case-based reasoning is also used extensively in day-to-day common-sense reasoning.

This model is called as R4 model of CBR. Because this model can be represented by schematic cycle which contains four Rs. Fig 2. depicts the R4 Cycle

a. Retrieve the most similar cases
b. Reuse the cases to attempt to solve the problem
c. Revise the proposed solution
d. Retain the new solution as a part of a new case

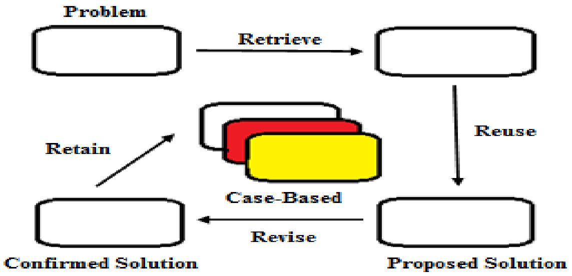


Figure 2. Case Based Reasoning R4 Cycle

ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network (ANN) is a computational model based on biological neural network. ANN also called Neural Network. It contains interconnected group of artificial neurons and processes the information by a connectionist approach.ANN is an adaptive system because it changes its structure based on information flow during the learning phase.

Basic topology of neural network consists of feed forward neural network and recurrent network. In feed forward neural network information flow starts from the input node. The information flow is one direction only from input node to hidden node and finally leads to the output node. In each node one or more processing elements (PE) may be active.PE is used to simulate the neurons in the brain.PE receive input from the outside world or from the previous layer. No cycles or loops in this network. But in recurrent neural network data flows bi-directionally and feedback connections exists here. Neural network consist of three parts architecture, learning algorithm and the activation function]. Neural networks are programmed to store, recognize and retrieve patterns or database entries for solving ill defined problems, to filter noise from measured data High Accuracy i.e. they are able to approximate complex non- linear mappings. They can be implemented in hardware. Ease of maintenance is another factor. When an element of the neural network fails it can continue without any problem [11]. They are independent from prior assumptions. Major application of Neural network are in Function Approximation, Classification, Data processing, Robotics. Fig 3. Shows the neural connectivity and Fig 4. Shows the Artificial neural network

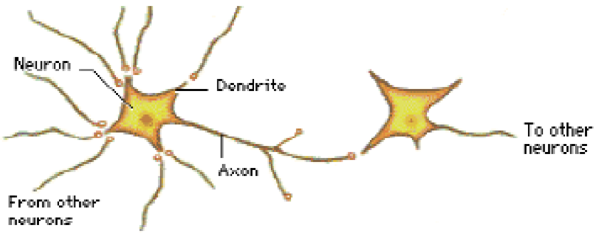


Figure.3 Neurons connectivity

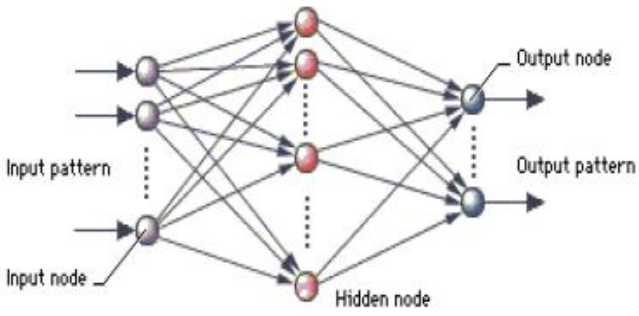


Figure 4. Artificial Neural Network

RESULT AND DISCUSSION

Based on the application of detecting seizure the various classification algorithms are tested for accuracy [9]. The Case Based Reasoning Classification results in 90.7% of specificity, 92.3% of sensitivity and 95.5% of prediction accuracy. Table 1. Shows the performance of the classification models. Fig 4. Shows the Comparison of the Classification Models

Table: 1

S8	Classification Algorithms	Sensitivity (%)	Specificity (%)	Accuracy(%)
1	K-Nearest Neighbor	67	76	72
2	Decision Tree	78	79	71
3	Bayesian network	87	84	85
4	Case Based Reasoning	90.7	92.3	95.5
5	Artificial Neural Networks	88	84	87

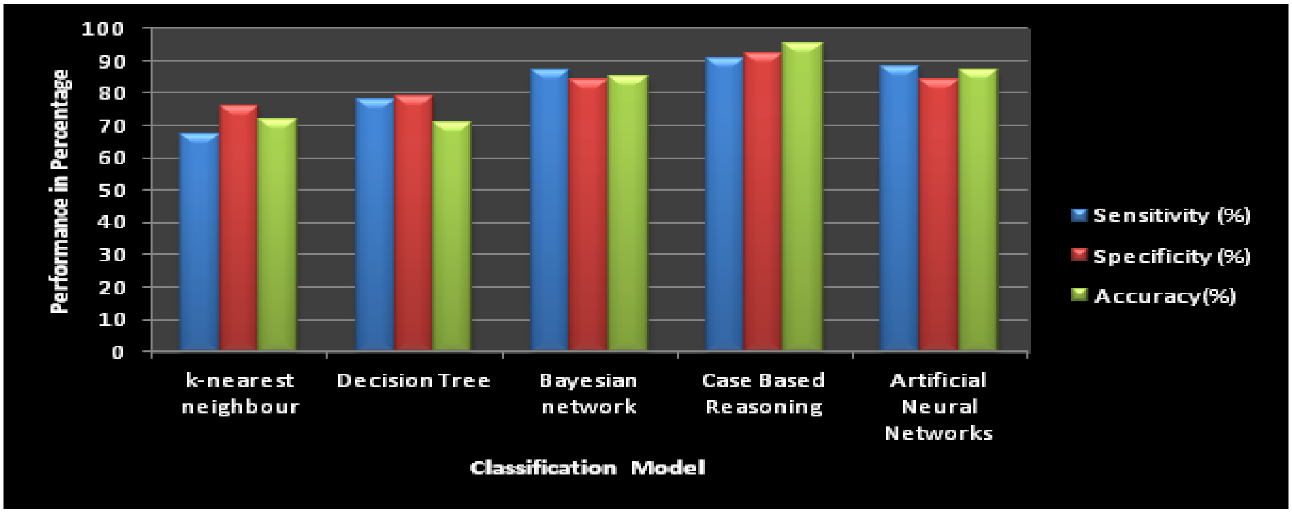


Figure 4. Comparison of the Classification Models

CONCLUSION

This paper deals with various classification techniques used in data mining and a study on each of them. Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. Each technique has got its own pros and cons as given in the paper. Based on the application of seizure the Case Based Reasoning shows the impressive accuracy.

REFERENCES

[1] Survey of Nearest Neighbor Techniques Nitin Bhatia (Corres. Author) Department of Computer Science DAV College Jalandhar, Vandana SSCS Deputy Commissioner's Office Jalandhar

[2] Survey of Classification Techniques in Data Mining: Thair Nu Phyu

[3] K-Nearest Neighbour Classifiers P´adraig Cunningham1 and Sarah Jane Delany2

[4] Decision Trees Andrew W. Moore Professor School of Computer Science Carnegie Mellon University

[5] A Fast Decision Tree Learning Algorithm Jiang Su and Harry Zhang Faculty of Computer Science University of New Brunswick, NB, Canada, E3B 5A3

[6] Top 10 algorithms in data mining XindongWu Vipin Kumar· J. Ross Quinlan · Joydeep Ghosh · Qiang Yang Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg © Springer-Verlag London Limited 2007

[7] K.-L. Tan, P.-K. Eng, and B.C. Ooi, "Efficient Progressive Skyline Computation," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2001.

[8] Charniak, E. 1991, .Bayesian Networks without tears. AI Magazine, Winter 1991.

[9] Ben-Gal I., Bayesian Networks, in Ruggeri F., Faltin F. & Kenett R Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007).

[10] Jordan, M.I. (1999). Learning in Graphical Models, MIT,Press, Cambridge.

[11] Ms. Aparna Raj, Mrs. Bincy G, Mrs. T.Mathu " Survey on common data mining classification Techniques", International Journal of Wisdom Based Computing , Vol 2, No.1,2012

Short Bio Data for the Author

Prof. P.Tamije Selvy received B.Tech (CSE), M.Tech (CSE) in 1996 and 1998 respectively from Pondicherry University. Since 1999, she has been working as faculty in reputed Engineering Colleges. At present , she is working as Assistant Professor(SG) in the department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore. She is currently pursuing Ph.D under Anna University, Chennai. Her Research interests include Image Processing, Data Mining and Pattern Recognition

Dr.V.Palanisamy received B.E (Electronics and communication), M.E (Communication Engineering) and PhD (Communication Engineering) in 1972, 1974 and 1987 respectively. Since 1974, he has been the faculty of Electronics and Communication Engineering and Served at various Government engineering colleges. At present, he is the principal at Info Institute of Engineering, Coimbatore. His research interest is in the Heuristic search methods for Optimization problems in various applications. He is a senior member of ISTE, SIEEE, and CSI.

Ms.S.Elakkiya has received Bachelor of Technology degree in Computer Science and Engineering under Anna University, Chennai in 2011. She is currently pursuing Master of Engineering degree in Computer Science and Engineering under Anna University, Coimbatore, India. Her areas of interest are image processing and data mining