International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), August 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SURVEY ON CLASSIFICATION ALGORITHMS FOR DATAMINING: (COMPARISON AND EVALUATION)

I.Bhuvana[1], Dr.C.Yamini[2]

[1]Research Scholar, [2]Associate Professor,Department of Computer Science,
Sri Ramakrishna college of Arts and Sciences for Women, Coimbatore, (India)

ABSTRACT

Data-Mining Classification is amachine learning technique used for portioning the data into different classes according to some constrains. It can deal with a wide variety of data so that large amount of data can be involved in processing. Several major kinds of classification methods that can be used such as decision tree induction, case-based reasoning, genetic algorithm, fuzzy logic techniques, C4.5, k-nearest neighbor classifier, Naive Bayes, SVM, and AdaBoost. In this paper a comparison among three classification's algorithms will be studied, these are (K- Nearest Neighbor classifier, Decision tree and Bayesian network) algorithms. The objective of this survey is to providean inclusive evaluationof different classification algorithms that are being generally used.

Index Terms: Bayesian network, Decision Tree, K-nearest neighbor classifier, KNN-Model.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the practice of analyzing data from different views and summarizing it into useful information. It is a wise technique that can be applied to extract useful patterns. In addition to collecting and managing of data, data mining also includes analysis and prediction.Exactly, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.Classification, Regression, Clustering, Rule generation, Discovering, association Rule…etc. each has its individualand different algorithms to attempt to fit a model to the data. Algorithm is a set of rules that must be followedwhen solving a specific problem (it is a finite sequence of computational steps that transform the given input toan output for a given problem).

Classification techniques in data mining are capable of processing anenormous 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 a certain part of data mining and is gaining more popularity [1].

In this paper Classification Method is considered, it focuses on a survey on various classification techniques that are most commonly used in data-mining. The study is a comparison between three algorithms (Bayesian network, K-NN classifier and Decision tree) to show the strength and accuracy of each algorithm for classification in term of performance efficiency and time complexity.

II. ALGORITHM

An algorithm usually means a small procedure that solves a recurrent problem.

A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations.A mapping from training sets to hypotheses that minimize the objective function.

III. ANALYSIS OF ALGORITHM

A situation may occur where many algorithms are available for solving a particular problem. The datastructure can be denoted in several ways and many algorithms are there to implement an operation on these data structure. Here to need a comparison of two algorithms to implement an operation on these data structure and the better one is chosen.

The study of an algorithm is mainly concentrate on time complexity and space complexity, as compared to time analysis the space analysis requirement for an algorithm is easer, but wherever necessary both of them are used.

The space refers to storage required to store the input data. The volume of memory needed by the program to run to completion isreferred to as Space complexity.The amount of time needed by the program to run to completion referred to as Time complexity, it is depending on the size of the input. It is a function of size: (n) [T (n)].

• Best Case:

It is the function defined by the maximum number of steps taken on any instance of size (n).

• Average Case:

It is the function defined by the Average number of steps taken on any instance of size (n).

• Worst Case:

It is the function defined by the minimum number of steps taken on any instance of size (n).

IV. K-NEAREST NEIGHBOR ALGORITHM

4.1 General View on KNN Algorithm

One of the simplest non parametric mechanism is used to identify the unknown data point based on the nearest neighbor whose value is already known. KNN algorithm is an easy to understand but has an incredible work in fields and practice specially in classification (it can be used in regression as well), non-parametric mean does not make assumptions on the data and that is great and useful in the real life, and lazy mean does not use training data to do generalization, that and in best case it makes decision based on the entire training data set.

For a data record t to be classified, its k nearest neighbors are retrieved, and this forms a neighborhood of record t. Majority voting among the data records in the neighborhood is usually used to decide the classification for record with or without attention of distance-based weighting . However, to apply KNN algorithm we need to choose an appropriate value for k, and the success of classification is dependent on the value of k. In a sense, the KNN method is predeterminedby k. There are many ways of choosing the K value, but a simple one is to run the algorithm many times with different k values and choose the one with the best performance [9].

There are three key elements:

• A set of labeled objects (e.g., a set of stored records)

• A distance or similarity metric to compute distance between objects.

• The value of *k*, the number of nearest neighbors [11].

KNN mainly works on the theory 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 [2]. Here a single number "k" 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 neighbor.

KNN requires:

- An integer k
- A training data set
- A metric to measure closeness

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions)

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Following fig.4.1 shows how the classification can be done based on the value of k. In the fig. there are three classes$\omega_1, \omega_2, \omega_3$ and we have to find a class label forx_u. Here suppose that the value of k is taken as 5 and calculate the Euclidian Distance between all current point and all other points. By examining we can find that of all the closest 5 points 4 of them belongs to ω_1 and one belong to ω_2 and hence$x_u$is assigned toω_1. The whole technique can be reviewed as determining the nearest neighbor and then finding its class using the neighbor values.

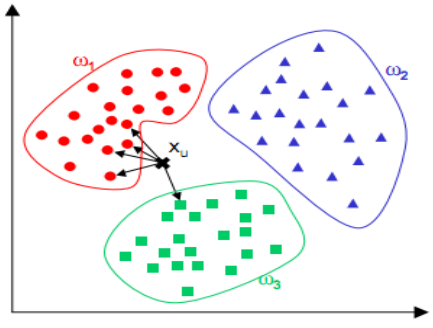

Fig. 4.1

Example of K-NN classification

Nearest neighbor classifiers are instance-based or lazy learners. They store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. This contrasts with eager learning methods, such a decision tree induction and back propagation, which construct a generalization model before receiving new samples to classify. Nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data.

4.2 Previously Researches on KNN Algorithm

Classification accuracy on six public datasets is comparable with C5.0,and KNN. KNN type classification method construct aKNN-modelwhich has a few representatives from training dataset with some extra information to represent the whole training dataset. The selection of each representation used the k, decided by dataset itself. The classification accuracy of KNN- Model was higher than KNN and C5.0. The KNN-Model significantly reduces the number of the data tuples in the final model for classification with a 90.41% reduction rate on average. It could be a good replacement for KNN in many applications such as dynamic web mining for a large repository[9]. A research over a medical data -set by[13] made a comparison between KNN and SVM, the result was after implementing these two algorithms, showed that KNN is a quite good classifier when applying KNN algorithm over small data set and the accuracy decrease when it applies over large data set. SVMis a complex classifier and the accuracy and other performance parameters are not too much depends over dataset size but about all factors dependent over the number of training cycles.The search time of SVM remains constant doesn't depend on the size of data set while search time in KNN increasing when the size of data increase [13]. Aresearch paper using cascading k-means clustering and KNN classifier over diabetic patient dataset and the result was quite good [10].

The model consists of three stages. The first stage, K-means clustering which is one of the simplest unsupervised learning algorithms and follows partitioning method for clustering. In the second stage Genetic algorithm (GA) and Correlation based feature selection (CFS) is used in a cascaded fashion in the third stage and a fine tuned classification is done using K-nearest neighbor (KNN) by taking the correct clustered instance of first stage and with feature subset identified in the second stage as inputs for the KNN. These stages enhanced classification accuracy of KNN. The proposed model obtained the classification accuracy of 96.68% for diabetic dataset [10]. Graz University of Technology, University of Washington, in May 2004 Experimented on the data of a surface inspection task and data sets from the UCI repository. Bayesian network classifiers more often achieve a better classification rate on different data sets as selective k-NN classifiers [14].A study on Classification Algorithms for Liver Disease Diagnosis results showed by [12] that the sensitivity of C4.5 classification algorithm and accuracy was less than KNN classifier accuracy and sensitivity.

Advantages of KNN Algorithm

• KNN is an easy to understand and easy to implement classification technique.

• It can perform well in many situations. Cover and Hart show that the error of the nearest neighbor rule isbounded above by twice the Bayes error under certain reasonable assumptions. Also, the error of thegeneral KNN method asymptotically approaches that of the Bayes error and can be used to approximate it.

• KNN is particularly well suited for multi-modal classes as well as applications in which an object canhave many class labels.

Disadvantages of KNN Algorithm

- It is easy to implement by computing the distances from the test sample to allstored vectors, but it is computationally intensive, especially when the size of the training set grows.

V. DECISION TREE ALGORITHM

Decision trees: Provide a graphical representation of a tree with conditions associated to the nodes that permit to classify a new instance in a predefined set of classes. DT have problems with very big data sets. It works with qualitative variables.

Decision Tree (DT) classification technique is the learning of decision trees from class labeled training tuples [2]. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. [2]. 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 [4].

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 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:

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

There are two possible types of divisions orpartitions:

- Nominal partitions: a nominal attribute may lead to a splitwith as many branches as values there are for the attribute.
- Numerical partitions: typically, they allow partitions like "X>" and "X <a". Partitions relating two different attributes are not permitted.
- What distinguish the different algorithms fromeach others are the partitions they allow, and whatcriteria they use to select the partitions.

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

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

Decision Tree can be considered as more interpretable connected to that of neural networks and support vector machines (SVM) 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 decisive the expected values from the end node back to the root node are known as decision tree roll-back.

Decision Tree shown in fig.5.1 and it 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 [6],[7]. Thus this can be applied to determine whether the climate suits well for playing golf. 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 severalapplications such as rolling a die, product decision, etc., for prediction analysis.

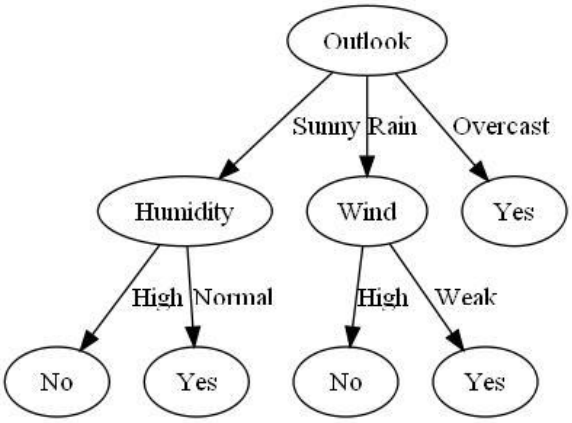

Fig. 5.1

Example for Decision Tree

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.

One major drawback of Greedy search is that it usually leads to sub-optimal solutions. A predictive model based on a branching series of Boolean tests.These smaller Boolean tests are less complex than a one-stage classifier.

(Entropy: a numerical measure of the uncertainty of an outcome)

Entropy of decision tree is the information gain measure, is minimized when all values of the target attribute are the same, If we know that commute time will always be short, then entropy = 0.

Entropy is maximized when there is an equal chance of all values for the target attribute (the result is random), If commute time = short in 3 instances, medium in 3 instances and long in 3 instances, entropy is maximized.

Calculation of entropy:

$(S) = \sum_{(i=1tol)} - |S_i|/|S| * \log_2(|S_i|/|S|)$

S = set of examples

Si = subset of S with value vi under the target attribute

l = size of the range of the target attribute.

Advantages of Decision Tree Algorithm

- Decision trees are simple to understand and interpret.
- They require little data and are able to handle both numerical and categorical data
- It is possible to validate a model using statistical tests.

- They are strong in nature,therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated
- Decision trees perform well with large data in a short time.
- Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.
- Nonlinear relationships between parameters do not affect tree performance
- The best feature of using trees for analytics - easy to interpret and explain to executives.

Disadvantages of Decision Tree Algorithm

- Decision-tree learners create over-complex trees that do not generalize the data well.
- Decision Trees do not work well if you have smooth boundaries. i.e. they work best when you have discontinuous piece wise constant model. If you truly have a linear target function decision trees are not the best.
- Decision Tree's do not work best if you have a lot of un-correlated variables. Decision tree's work by finding the interactions between variables. If you have a situation where there are no interactions between variables linear approaches might be the best.
- Data fragmentation: Each split in a tree leads to a reduced dataset under consideration. And, hence the model created at the split will potentially introduce bias.
- High variance and unstable : As a result of the greedy strategy applied by decision tree's variance in finding the right starting point of the tree can greatly impact the final result. i.e. small changes early on can have big impacts later. So- if for example you draw two different samples from your universe, the starting points for both the samples could be very different (and may even be different variables) this can lead to totally different results.

VI. BAYESIAN NETWORK

Bayesian networks are a statistical method for Data Mining, a statistical method for discovering valid, novel and potentially useful patterns in data. Bayesian network (BN) is also called belief networks, is a graphical model for probability relationships among a set of variables features. BN consist of two components:

- First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables.
- Second component is a set of parameters that describe the conditional probability of each variable given its parents. A Bayesian network (BN) describes a system by specifying relationships of conditional dependence between its variables .The conditional dependences are represented by a directed acyclic graph, in which, each node [8].

A Bayesian network specifies a joint distribution in a structured form. Represent dependence/independence via a directed graph.

- Nodes = random variables
- Edges = direct dependence

- Structure of the graph ⇔ Conditional independence relations.

  In general, Bayesian network is expressed as

  p(X1, X2,....XN) = Π p(Xi | parents(Xi))

  ↑

  The full joint distribution ↑

The graph-structured approximation

- •Requires that graph is acyclic (no cycle)
- Two components to a Bayesian network: The graph structure (conditional independence assumptions)

The numerical probabilities (for each variable given its parents)There are several equivalent definitions of a Bayesian network.For all the following, let *G = (V, E)* be a directed acyclic graph (orDAG), and let X = (Xv)v ∈□ V be a set of randbm variables indexed by V.

A Bayesian network is a DAG, G= (V, E) and a set of conditional probability distributions P. Each node has conditional probability table (CPT) which quantifies the effect of parent node.

BNs take account of prior information for a given problem. This prior expertise about the structure of Bayesian network can take the following forms:

- Declare that a node is root node.
- Declare that a node is leaf node.
- Declaring that a node has direct effect of another node.
- Declaring that a node is not directly connected to another node.
- Declaring that two nodes are independent, giving a condition set.
- Providing partial ordering among the nodes.

Main advantages of Bayesian Network.

- Bayesian Network can be used by investigators to use their domain expert knowledge in the knowledge discovery process but other techniques primarily depend upon coded data to extract knowledge.
- BN model can be easily understood compared to many other techniques by the use of nodes and arrows. Researchers can encode the domain expert knowledge by the use the graphical diagrams, so they can easily understand the output of BN.
- Bayesian Network supports the use of probabilistic inference to update and revise belief values.
- Bayesian networks readily permit qualitative inferences without the computational inefficiencies of traditional joint probability determinations. In doing so, they support complex inference modeling including rational decision making systems, value of information and sensitivity analysis.

As such, they are useful for causality analysis and through statistical induction they support a form of automated learning.

Applications of Bayesian Network are finding Relative Military Strength, River Crossing under Fire, Enemy Intention and Medical Diagnosis.

Disadvantages of Bayesian Network.

- Experts may be challenged to express their knowledge in the form of probability distributions

- Some BN software packages may have limited ability to deal with continuous data.
- The acyclic property BN is required to carry out probability calculus, but implies that feedback effects cannot be included inthe network.

Table 1 Comparision of Classification Techniques

Method	**Generative Or Discriminative**	**Loss Function**	**Parameter estimation algorithm**
K-Nearest Neighbor	Discriminative	-log(X,Y) Or Zero-one loss	Should store all training data to classify new points
Decision Tree	Discriminative	Zero-one loss	C4.5

VII. CONCLUSION

Due to our survey on comparison among data mining classification's algorithms (Decision tree, KNN,Bayesian) and analyzing of the time complexity of the stated algorithms we determine that all decision Tree's algorithms have less error rate and it is the easier algorithm as compared to KNN and Bayesian algorithms. The knowledge in Decision Tree represented in form of [IF-THEN] rules which is easyto understand. The disadvantages of decision tree algorithm are usuallyrequiring certain knowledge statistical experience. This leads to complete the process accurately. It can also be difficult to include variables on the decision tree, exclude duplicate information. As we mentioned there are many specific decision-tree algorithms. CART decision tree algorithm is the best algorithm for classification of data, which have shortest execution time. The result to predictive data mining technique on the same dataset showed that Decision Tree outperforms and Bayesian classification having the same accuracy as of decision tree but other predictive methods like KNN, Classification based on clustering are not giving good results. Due to our survey based on the previously researches we extract the fact that among (Decision tree, KNN, Bayesian) algorithms in data mining, KNN is having lesser accuracy while Decision tree and Bayesian are equal. But if Decision tree algorithm has merged with genetic algorithm then the accuracy of the Decision tree algorithm will improve and become more powerful and it will arise to be the best model approach among the other two algorithms. The efficiency of results using KNN can be improved by increasing the number of data sets and for Bayesian algorithm classifier by increasing the attributes. For time issue, researches statistics we conclude thatthe faster algorithm for classifier respectively is: Navi- Bayes algorithm, Decision tree and finally KNNalgorithmthat mean the last one is the most slowly algorithm for classifier.

REFERENCES

[1] RAJ, M. A. 2012. Mrs. Bincy G, Mrs. T. Mathu. Survey on common data mining classification Technique. International Journal of Wisdom Based Computing, 2.

[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 · JoydeepGhosh · 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] BAYAT, S., CUGGIA, M., ROSSILLE, D., KESSLER, M. & FRIMAT, L. Year. Comparison of Bayesian Network and Decision Tree Methods for Predicting Access to the Renal Transplant Waiting List. In: MIE, 2009. 600-604.

[9] GUO, G., WANG, H., BELL, D., BI, Y. & GREER, K. 2003. KNN model-based approach in classification. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer.

[10] KAREGOWDA, A. G., JAYARAM, M. & MANJUNATH, A. 2012. Cascading K-means Clustering and KNearest Neighbor Classifier for Categorization of Diabetic Patients. International Journal of Engineering and Advanced Technology (IJEAT) ISSN**,** 2249-8958.

[11] WU, X., KUMAR, V., QUINLAN, J. R., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G. J., NG, A., LIU, B. & PHILIP, S. Y. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems, 14**,** 1-37.

[12] RAMANA, B. V., BABU, M. S. P. & VENKATESWARLU, N. 2011. A critical study of selected classificationalgorithms for liver disease diagnosis. International Journal of Database Management Systems, 3**,** 101-114.

[13] RAIKWAL, J. & SAXENA, K. 2012. Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set. International Journal of Computer Applications, 50, 35-39.

[14] PERNKOPF, F. 2005. Bayesian network classifiers versus selective k-NN classifier. Pattern Recognition, 38**,** 1- 10.