Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.4, No.8, 2013
www.iiste.org

Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)

Delveen Luqman Abd AL-Nabi[1*] Shereen Shukri Ahmed[2]

1. School of Business Management, Duhok University, Zakho Street 38, 1006AJ Duhok, Kurdistan Region-Iraq.
2. College of Computer Science, Zakho University, Zakho Road, Kurdistan Region-Iraq.

* dlvin_lukman@yahoo.com

Abstract
Data mining concept is growing fast in popularity, it is a technology that involving methods at the intersection of (Artificial intelligent, Machine learning, Statistics and database system), the main goal of data mining process is to extract information from a large data into form which could be understandable for further use. Some algorithms of data mining are used to give solutions to classification problems in database.

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 paper will demonstrate the strength and accuracy of each algorithm for classification in term of performance efficiency and time complexity required. For model validation purpose, twenty-four-month data analysis is conducted on a mock-up basis.
Keywords: Decision tree, Bayesian network, k- nearest neighbour classifier.

1. Introduction

Data Mining (The analysis step of the knowledge discovery in data base) a powerful new technology improved and so fast grown. It is a technology used with great potential to help business and companies focus on the most important information of the data that they have to collect to find out their customer's behaviors. Intelligent methods are applied in order to extracting data pattern, by many stages like" data selection, cleaning, data integration, transformation and pattern extraction". Many methods are used for extraction data like" Classification, Regression, Clustering, Rule generation, Discovering, association Rule…etc. each has its own and different algorithms to attempt to fit a model to the data. Algorithm is a set of rules that must be followed when solving a specific problem (it is a finite sequence of computational steps that transform the given input to an output for a given problem). The problem can be a machine.

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 (RAJ *et al*. 2012)

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. Next section deals with a study on Algorithm, section III describe what algorithm analysis is and what time is and space complexity, in section IV k-nearest neighbor mechanism has explained. Section V describes Decision Tree and section VI deals with Bayesian network, finally last section concludes the paper.

2. Algorithm

The algorithm is a computational procedure which takes some value or set of value as input and generates some value or set as output. The result of a given problem is the output that we got after solving the problem. The algorithm is considered to be correct , if for every input instance, it generate the correct output and it gets terminated and give the desired output otherwise it does not considered as a correct algorithm.

3. Analysis of Algorithm

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

The analysis of algorithm is focus 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 in addition to the space required to store the input data. The amount of memory needed by the program to run to completion referred 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).

4. K-Nearest Neighbour Algorithm

4.1 General view on KNN Algorithm

One of the simplest non parametric lazy algorithms called 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. 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. Figure 1 illustrates the modeling.

For a data record t to be classified, its k nearest neighbors are retrieved, and this forms a neighborhood of t. Majority voting among the data records in the neighborhood is usually used to decide the classification for t with or without consideration of distance-based weighting . However, to apply KNN algorithm we need to choose an appropriate value for k, and the success of classification is very much dependent on this value. In a sense, the KNN method is biased by 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 (GUO *et al*. 2003)

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. (WU, KUMAR *et al*. 2008)

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 is bounded above by twice the Bayes error under certain reasonable assumptions. Also, the error of the general 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 can have many class labels.

Disadvantages of KNN Algorithm:

The naive version of the algorithm is easy to implement by computing the distances from the test sample to all stored vectors, but it is computationally intensive, especially when the size of the training set grows.

4.2 Previously researches on KNN algorithm

A group of researchers in University of Ulster and Queen's University Belfast in them research find the classification accuracy on six public datasets is comparable with C5.0 and KNN and a novel KNN and they named it KNN model which has a few representatives from training dataset with some extra information to represent the whole training dataset the selection of each representative they used the optimal but different 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 (GUO *et al*. 2003). In (RAIKWAL, J. & SAXENA, K. 2012) did a research over a medical data set they made a comparison between KNN and SVM them result was after implementing the two algorithm showed that K-NN is a quit good classifier but when applying KNN algorithm over small data set and it is accuracy decrease when it applies over large data set it performs poor results (it's all performance parameters are varies according to the size of dataset)..SVM is complex classifier and the accuracy and other performance parameters are not too much depends over dataset size but about all factors dependent over the no 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(RAIKWAL, J. & SAXENA, K. 2012). (KAREGOWDA, A. G., JAYARAM, M. & MANJUNATH, A. 2012) made a paper using cascading k-means clustering and KNN classifier over diabetic patient them result was quite good.

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 a fine tuned classification is done using K-nearest neighbor (KNN) by taking the correctly clustered instance of first stage and with feature subset identified in the second stage as inputs for the KNN. These enhanced classification accuracy of KNN. The proposed model obtained the classification accuracy of 96.68% for diabetic dataset (KAREGOWDA *et al*. 2012). 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 (PERNKOPF, F. 2005). (VENKATESWARLU *et al*. 2011) made a study on Classification Algorithms for Liver Disease Diagnosis them results showed that the sensitivity of C4.5 classification algorithm and accuracy was less than KNN classifier accuracy and sensitivity (RAMANA
& VENKATESWARLU *et al*. 2011)

5. Decision Tree Algorithm

5.1 General idea of algorithm

Tree structure which has been widely used to represent classification models (a classifier depicted in a flowchart) (BARROS *et al*. 2012). Decision tree induction algorithms, an inductive learning task use particular facts to make more generalized conclusions. Most decision tree induction algorithms are based on a greedy top-down recursive partitioning strategy for tree growth. They use different variants of impurity measures, like; information gain (BARROS *et al*. 2012), gain ratio (WANG *et al*. 2005), and distance-based measures (DE MÁNTARAS & R. L. 1991), to select an input attribute to be associated with an internal node. 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. The general form of this modeling approach is illustrated in Figure 2.
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:

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

- S = set of examples
- Si = subset of S with value vi under the target attribute
- l = size of the range of the target attribute.

Decision Trees offer many benefits in data mining technology like:

- Self-explanatory and easy to follow when compacted
- The ability of handling variety of input data: nominal, numeric and textual
- Ability of processing data sets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Useful for various tasks, such as classification, regression, clustering and feature selection

Decision tree classifier is able to break down a complex decision making process into collection of simpler and easy decision. The complex decision is subdivided into simpler decision. It divides whole training set into smaller subsets. Information gain, gain ratio, gain index are three basic splitting criteria to select attribute as a splitting point. Decision trees can be built from historical data they are often used for explanatory analysis as well as a form of supervision learning. The algorithm is design in such a way that it works on all the data that is available and as perfect as possible (PAWAR, T. & KAMALAPUR. 2011). There are many specific decision-tree algorithms:

1. ID3 (Iterative Dichotomiser 3)
2. C4.5 Algorithm, Successor of ID3
3. CART (Classification And Regression Tree)
4. MARS: extend decision tree to better handle numerical data.

Advantage Of Decision Tree:Advantages over other learning algorithms, such as robustness to noise, low computational cost for generating the model, and ability to deal with redundant attributes, Besides, the induced

model usually presents a good generalization ability (HAN *et al*. 2006),(PANG-NING *et al*. 2006)
Problems with Decision Tree:

- While decision trees classify quickly, the time for building a tree may be higher than another type of classifier
- Decision trees suffer from a problem of errors propagating throughout a tree a very serious problem as the number of classes' increases.

5.2 Different Between Nearest Neighbor Classifiers and Decision Tree Algorithm

Nearest neighbor classifiers are instance-based or lazy learners in that 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. Lazy learners can incur expensive computational costs when the number of potential neighbors (i.e., stored training samples) with which to compare a given unlabeled sample is great. Therefore, they require efficient indexing techniques. An expected lazy learning method is faster data training than eager methods, but slower at classification since all computation is delayed to that time. Unlike decision tree induction and back propagation, nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data.

5.3 Previously researches of Decision Tree

(Fahad Shahbaz Khan *et al*. 2008), they did an experiment to examine ID3 and C4.5 in oral medicine after the experiment they selected the C4.5 decision tree algorithm because the algorithm has the ability for handling data with missing attribute values better than ID3 decision tree algorithm. It also avoids overfitting the data and reduces error pruning (KHAN *et al*. 2008). 2008 Author Patrick Ozer made acomparison among four classification decision tree algorithm J84, REPTree, RandomTree and LMT the J48 algorithm is WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for 20 classification and uses reduced-error pruning REPTree is a fast decision tree learner. RandomTree is an algorithm for constructing a tree that considers K random features at each node. Performs no pruning.LMT is a combination of induction trees and logistic regression. LMT uses cost-complexity pruning The LMT algorithm seems to perform better on data sets with many numerical attributes. The results showed that LMT has got the best overall performance, the performance of the J48 and RepTree algorithm were almost the same on all the data sets, RandomTree algorithm we see that it builds the largest trees (and has the lowest overall performance). For the LMT algorithm restricted the experiments done on five runs, because of the time it costs to run that algorithm. For the other three algorithms one run took less than a minute, but for the LMT algorithm one run could take several hours. This algorithm is significantly slower than the other algorithms, for the four classification tree algorithms they used, that using cost-complexity pruning has a better performance than reduced-error pruning (OZER & P. 2008). (Rahul *et al*. 2012) modified ID3 decision tree algorithm and named it improved ID3, After testing the original ID3 algorithm and proposed improved ID3 algorithm on dataset. In improved ID3 algorithm they got more number of nodes and more number of rules which means that improved ID3 algorithm is more efficient than original ID3 algorithm. And it is differs from original ID3 in following way.

- It is using extra Association Function to overcome the short comings of Id3
- Improved Id3 more reasonable and effective rules are generated
- Missing values can be considered and will not have impact on accuracy of decision.
- Accurate rules
- The accuracy of decision is more because e no of rules is more

The only Disadvantages were Time complexity is more in improved ID3, but it can be neglected because now day's faster and faster computers are available (Rahul *et al*. 2012).
October 2012 College Of Engineering ,Pune, India present a paper which compare between ID3 decision tree and FID3(Fuzzy iterative dichotomizer3) which is same to ID3 but add fuzzification for improving the result of ID3 , the data represent in ID3 is crisp while for FID3 they are fuzzy, with continuous attributes the main difference between ID3 is that ID3 works well with discrete values but for continues value FID3 showed better accuracy .Classical decision tree has two issues like How to split the training instances and what the stopping rule is to terminate the splitting procedure. These two problems are solved with the help of fuzzy decision tree. (SURYAWANSHI, R. D. & THAKORE, D. 2012) a group of researchers in France did a comparison between Bayesian network algorithm and CART decision tree for predicting access to the renal transplant waiting list them results showed the models were complementary sensitivity, specificity and positive predictive of the two models was same result and both model had high accuracy since the Bayesian network provided a global view of the variables' associations while the decision tree was more easily interpretable by physicians (BAYAT *et al*. 2009).

6. Bayesian Network Algorithm

6.1 General view of the algorithm

Bayesian network (BN) is also called belief networks, is a graphical model for probability relationships among a set of variables features, This 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 (BAYAT *et al*. 2009).The general form of this modeling approach is illustrated in Figure 3.

6.2 Previously Researches Of The Algorithm

March 2011 group of researchers did a comparison among many classification algorithm in data mining over Heart Disease Prediction , in them comparison the result showed that the accuracy of Navi Bayesian and decision tree is so close to each other we can say that both of them have same accuracy . while the accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or other features, and the time taken by each algorithm showed that the Bayesian algorithm was the faster one ,but by incorporated genetic algorithm with decision tree algorithm in this way Decision tree algorithm will be outperform on KNN and Bayesian algorithm in the manner of accuracy(SONI *et al*. 2011). In many data mining classification model the decision tree and Bayesian algorithm had similar high predictive performance Bayesian networks can link more variable in complex direct and indirect ways making interpretation more complex while decision trees and provide a simpler and more direct interpretation (BAYAT *et al*. 2009). International Journal on Computer Science and Engineering (IJCSE) J.Sreemathy Research Scholar Karpagam University and P. S. Balamurugan Research Scholar ANNA UNIVERSITY published a paper which was about a comparison between KNN and Bayesian classification algorithm on an efficient text classification the comparison showed the Precision of Bayesian algorithm over KNN and SVM algorithm (SREEMATHY *et al*. 2012).the success and outperform of Bayesian algorithm over KNN appear in another set of data base which is Tuberculosis . Classify the patient affected by tuberculosis into two categories (least probable and most probable) (Hardik Maniya *et al*. 2011). Comparison four different data mining-based techniques incl
uding artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs) and Bayesian networks (BNs) were investigated. Results of validation stage showed ANN had the highest classification accuracy, 96.33%. After ANN, SVM with polynomial kernel function (95.67%), DT with J48 algorithm (94.67%) and BN with simulated annealing learning (94.33%) had higher accuracy, respectively (MOLLAZADE *et al*. 2012).

7. 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 mentioned algorithms we conclude that all decision Tree's algorithms have less error rate and it is the easier algorithm as compared to KNN and Bayesian. The knowledge in Decision Tree represented in form of [IF-THEN] rules which is easier for humans understand.The disadvantages of decision tree algorithm are typically require certain knowledge statistical experience 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, Neural Networks, Classification based on clustering are not giving good results. up to here and 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 in this way 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 that the faster algorithm for classifier respectively is: Navi- Bayes algorithm, Decision tree and finally KNN algorithm that mean the last one is the most slowly algorithm for classifier.

References

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

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.
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.
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.
KAREGOWDA, A. G., JAYARAM, M. & MANJUNATH, A. 2012. Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. *International Journal of Engineering and Advanced Technology (IJEAT) ISSN***,** 2249-8958.
PERNKOPF, F. 2005. Bayesian network classifiers versus selective k-NN classifier. *Pattern Recognition,* 38**,** 1-10.
RAMANA, B. V., BABU, M. S. P. & VENKATESWARLU, N. 2011. A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems,* 3**,** 101-114.
BARROS, R. C., BASGALUPP, M. P., DE CARVALHO, A. & FREITAS, A. A. 2012. A survey of evolutionary algorithms for decision-tree induction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 42**,** 291-312.
WANG, Y., MAKEDON, F. S., FORD, J. C. & PEARLMAN, J. 2005. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics,* 21**,** 1530-1537.
DE MÁNTARAS, R. L. 1991. A distance-based attribute selection measure for decision tree induction. *Machine learning,* 6**,** 81-92.
PAWAR, T. & KAMALAPUR, S. A Survey on Privacy Preserving Decision Tree Classifier.
HAN, J., KAMBER, M. & PEI, J. 2006. *Data mining: concepts and techniques*, Morgan kaufmann.
PANG-NING, T., STEINBACH, M. & KUMAR, V. Year. Introduction to data mining. *In:* Library of Congress, 2006.
KHAN, F. S., ANWER, R. M., TORGERSSON, O. & FALKMAN, G. 2008. Data mining in oral medicine using decision trees. *World Academy of Science, Engineering and Technology,* 37**,** 225-230.
OZER, P. 2008. Data Mining Algorithms for Classification.
PATIL, R. A., AHIRE, P. G., PATIL, P. D. & GOLANDE, A. L. A Modified Approach to Construct Decision Tree in Data Mining Classification.
SURYAWANSHI, R. D. & THAKORE, D. 2012. Decision Tree Classification Implementation with Fuzzy Logic. *IJCSNS,* 12**,** 93.
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.
SONI, J., ANSARI, U., SHARMA, D. & SONI, S. 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications,* 17**,** 43-48.
SREEMATHY, J. & BALAMURUGAN, P. 2012. An efficient text classification using knn and naive bayesian. *International Journal on Computer Science and Engineering,* 4**,** 392-396.
Hardik Maniya , Mohsin Hasan, Komal Patil. 2011. Comparative study of Naïve Bayes and KNN for Tuberculosis. *International Journal of Computer Applications(JCA*),2011.
MOLLAZADE, K., OMID, M. & AREFI, A. 2012. Comparing data mining classifiers for grading raisins based on visual features. *Computers and Electronics in Agriculture,* 84**,** 124-131.

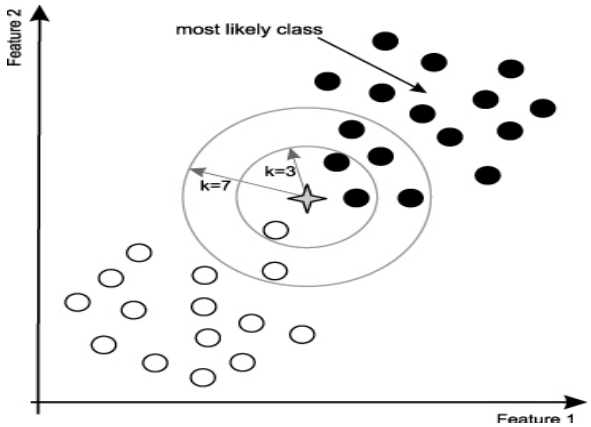

Figure 1. Illustration of K-Nearest Neighbors.

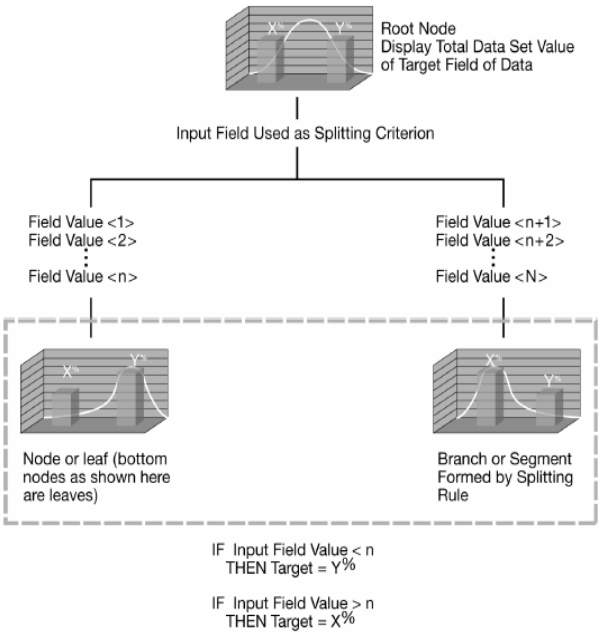

Figure 2. Illustration of Decision Tree.

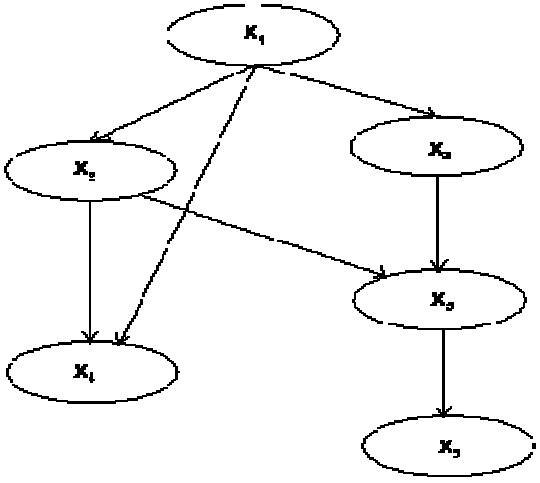

Figure 3. Illustration of Bayesian Network algorithm.

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

CALL FOR PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar