RESEARCHERID
THOMSON REUTERS
[COTII- 2018]
ISSN 2348 – 8034
Impact Factor- 5.070
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

DATA MINING CLASSIFICATION TECHNIQUES: A SURVEY

Rajat Verma1, Dr. Namrata Dhanda[2], Ms. Shikha Singh[3] & Harshita Mishra[4]
[1&4]M.Tech, Computer Science & Engineering, Amity School of Engineering & Technology, Amity University, Lucknow
[2]Professor, Department of Computer Science, Amity School of Engineering & Technology, Amity University, Lucknow
[3]Assistant Professor, Department of Computer Science, Amity School of Engineering & Technology, Amity University, Lucknow

ABSTRACT

When it comes to classification in concern to data mining and machine learning, it is considered as the most important. Classification techniques are well suited for the enormous data that are considered as unprocessed raw facts and figures for some people and when it is processed it is termed as information for that same group of individuals. As it has the inference rules from the beginning its work is to categorize the data and provide a class label and all the work should be done in an organized way. Classifying the freshly present unprocessed raw facts and figures into a label that determines the class is basically the work of classification models. It focuses on searching a hypothesis in terms of machine learning that illustrates the concept of the data and that allows it to differ from other classes. It deals with basically discrete values in the case of supervised learning as when it comes to continuous data, regression is followed. Making an exact classifier that is fast efficient is a pre-requisite in the mining process and the discovery of the proficiency when it deals with a concept known as knowledge that corresponds to the 5th and 7th process in the entire process of Knowledge discovery in data base abbreviated as KDD. Prediction having a classification approach is done and thus resulting in determining the labels of the class and that too is based on a category, and training set or inference rules does the work of classifying. It is a two-step process. A study has been done in this paper focusing on the research background of various data mining techniques approaching classification. Few examples of them includes:

- ✓ *Decision Tree*
- ✓ *K-Nearest Neighbor*
- ✓ *Support Vector Machines*
- ✓ *Naive Bayesian Classifiers*
- ✓ *Neural Networks*

Keywords: *Prediction, Classification, Model, Categories.*

I. INTRODUCTION

Classification is a two-step process. Initial step includes that a model (hypothesis) that has to be madedepending on a few number of training data set (inference rules), whereas the seconds step uses that hypothesis to classify/categorize an unknown tuple/dataset into a name considered as a label denoting a class.

Discrete data is considered with the Classification process, when it deals with continuous data regression is followed. A diagrammatic/pictorial representation is given concerning the construction of a model.

ISSN 2348 – 8034
Impact Factor- 5.070

1.1 Step-1:Model Construction: Initial Step

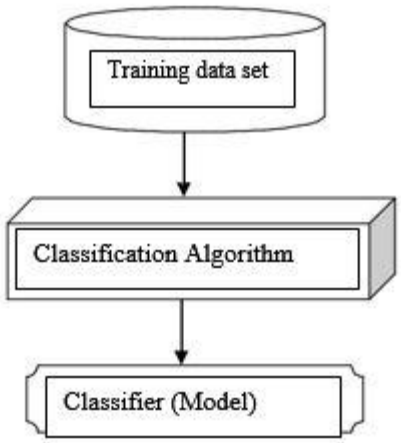

Fig.1 - Model construction step

The Second step comprises of selection of unknown tuple that can be placed in the same category. In simple words the use of classifier is done and it's shown with the help of a figure below.

1.2 Step 2 - Model used for unknown tuple

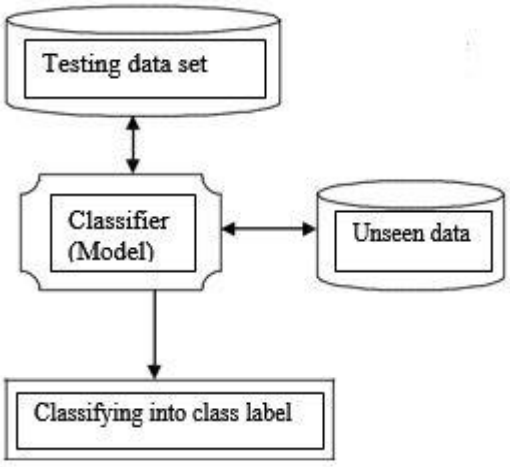

Fig. 2 Use of Classifier

ISSN 2348 – 8034
Impact Factor- 5.070

II. CLASSIFIERS CHARACTERISTICS

Various kinds of classifiers are present and each of them differs from the other because of some uniqueness. The traits are termed as classifier's characteristics. These characteristics are illustrated below:

- **Correctness:** -Accuracy is followed till which extent. Some numerical values can be used and based on number of tuple, identification can done as correct or wrong.
- **Time:** - Calculating the time that is needed to make or finalize the model? , indicating its time complexity. Prediction time as well as computational costs are calculated.
- **Strength:** -Real world data is full of noise and inconsistency even though if this provide classification of tuple it indicates that it has strength. Wrong values or missing values can be considered as noise and will be removed via smoothening.
- **Data Size:** -The size of the database should be kept independent from the classifier. The scalability feature should be present in the hypothesis. The performance measure is independent to the database's immensity.
- **Extendibility:** - Whenever it is required, additional or new features can be added. This is somehow complex as one needs to know the entire thing before adding something.

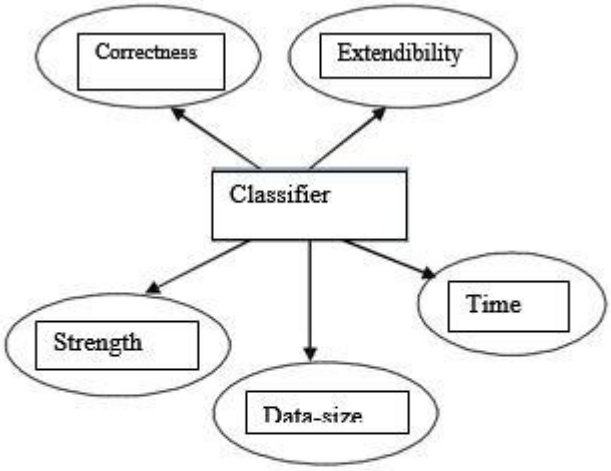

Fig. 3 Characteristics of Classifier

III. RESEARCH BACKGROUND

Associative Classification and Genetic Algorithm for Heart disease was proposed byAkhilJabbar et al. in 2012. Prediction of heart disease that uses genetic approach was proposed and it too used the associative classification algorithm. The pros of this algorithm is to uncover the rule of prediction, high prediction accuracy and high level of interestingness.Genetic algorithm approach is also helpful for medical scenario in prediction of heart disease and in making diagnosis decision appropriately.[1].

Categorization of Heart Diseases by Artificial Neural Network probably abbreviated as ANN and Feature Subset Selection was proposed by AkhilJabbar et al.in 2013. For categorization of heart disease new approach was proposed.It was known as new feature selection method by ANN. In this method attributes were ranked that divided

heart diseases into different features and this reduced the number of diagnosis test, to be taken by patients.By this proposed method, ineffective data and steps were removed. [2].

For medical diagnosis, Gain ratio method was proposed by N. S. Nithya et al. in 2014 which is based on mining classifier using weighted fuzzy association rule algorithm (fuzzy logics). In the proposed paper it is shown that earlier method i.e. algorithm like fuzzy association rule mining and information gain, for extracting association rule and membership function was not realizable/appropriate in some situation therefore not to considered them as a feasible algorithm. In proposed method huge number of distinct values are used and gain ratio was improved peculiarly that was based on weighted fuzzy association rule mining.[3].

S. OlalekanAkinola and O. JephtharOyabugbe proposed "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" in the year 2015. The study that was proposed by them was designed for the determining that how data mining algorithm having a classification approach performs with the increment in concern to input data sizes.3 data mining algorithms having a discrete approach were used in this scenario.

When subjection is done in concern to the data immensity but is varying in nature is concerned then the following algorithms plays a vital role. Some of them are –

- ✓ Decision Tree
- ✓ Multi-Layer Perceptron (MLP) Neural Network
- ✓ Naïve Bayes.

The time complexities occupied by these finite procedures for the intention of training and maintaining the efficiencies in terms of correctnessconcerning classifications were seen with focused attention for the varying sizes of data.

For prediction of cancer, Data Mining Technique has been considered and the researchers are in [5], this happened in the year 2015.Prediction of cancer was done by the use of data mining technique. Classification and Association in data mining was used for diagnosis of cancer.The diagnosis cancer in benign and malignantpatients, they used FP algorithm in Association Rule that generated result in the form of pattern.

By using Data mining Classification method, a management system known as the Appraisal Management System, was proposed by Nikhil N. Salvithal& R.B. Kulkarni in 2016. Assorted classifier algorithm was proposed that was applied on Talent data set to find talent data set and judge their performance. Count on accuracy is one of the best classifier that is used to make protocols for approach having classification to find whether potential talents provides benefits or vice-versa.[6].

Performance analysis concerning Classification in respect to mining of data technique was proposed by Tanvi Sharma &Anand Sharma in 2016 that was based on healthcare in a public concern.Application of classification in respect to mining of data using preprocessing tools, example of it are WEKA as well as rapid miner were preferred for determining the system of health care in comparison to public health dataset. To measure the standard of performance measure, the accuracy percentage of every data mining is used. The technique with best accuracy is chosen as best technique for a data set[7].

IV. TYPES OF CLASSIFICATION MODELS

For the process ofmaximization of the anticipatingcorrectness that is realized by the different models, classification approaches are preferred. Classification task is known as supervised technique in which class consist of instance. Several models are present that are used for classification- [8, 9and 10].

- ✓ Decision Tree,
- ✓ K-Nearest Neighbor,

ISSN 2348 – 8034
Impact Factor- 5.070

- ✓ Bayesian Classifier
- ✓ Neural Networks
- ✓ Support Vector Machines

4. A: Decision Trees

Ituses tree like structure and uses recursive partition (partitioning again and again) of the possible distinct instances. Itis a classification mechanism. Nodes and root are present in decision tree model. Other than root, nodes can multiple outgoing edges but can have one edge coming towards it.

A test is performed on intermediate nodes or test nodes and then they make an edge that is going outwards. Nodes that do not have outgoing edges are called leaves or terminal nodes or decision node. In this mechanism, each immediate node bifurcates/divides the instance space into two sections or possibly smaller spaces and a function having discrete characteristic concerning values of input is maintained.

Instance space is divided into sub spaces which is having a discrete approach/function to the input attributes value by internal node in decision tree.

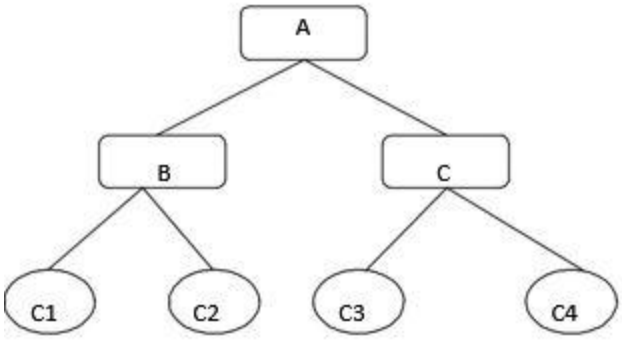

Fig.4- Decision Tree Classifiers

A mechanism of decision tree classifier is shown in the above figure, showing bifurcation at every node except at the leaves.

4. B:K-Nearest neighbor

Samples are trained by learning process in K-Nearest neighbor. In n-dimensional space, samples are represented as points and samples concern to training are gathered in n-dimensional pattern space. When the case of unknown sample is progressed, a k-nearest neighbor classifier does the finding procedurein the spaceof pattern for the random number of training samples (say k)that having the closest measure to the sample that is not known. Euclidean distance that is a distance measure, is used to define "Closeness" between two points.

To find Euclidean distance between two points i.e. X=(x1, x2… xn) and Y= (y1, y2… yn) is given as d(X, Y).

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

ISSN 2348 – 8034
Impact Factor- 5.070

Equal weight are assigned to each attribute in the case of nearest neighbor classifier. A major application of this is prediction, indicating that to return a prediction which is real-valued for a given sample that is unknown.

4. C:Bayesian Classifiers

Their alias can be statistical classifiers. Probability is used for predicting class membership. Inputs with high dimensionality are used in the Naïve Bayes Classifier. Naive Bayes can outshine more worldlythan other classification methods.

Let a training set named as Dwith associated class labels. N-dimensional attributes represents tuples namely A1, A2… An. Let us assume that there are l classes represented by f variable, f1, f2… fl. Suppose a tuple is given named X, the classifying mechanism or basically called as the classifier will tell that the tuple X belongs to that class which would be having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier determines that tuple x belongs (∈) to the class named as fi iff the following condition is valid. The condition is illustrated as: P (fi / X) > P (fj/X) for 1<= j <= l, j≠i. Thus we do the process of maximizationin concern to P(fi / X). The class fi for which P(fi / X) is maximized ca be termed as maximum posteriori hypothesis. According to the Bayes' theorem-

$$P(fj/X) = \frac{P(X/fi)\,P(fi)}{P(X)}$$

P(X) is a constant scenario universally for all the classes, only P (X/fi) P (fi) need be maximized that is the numerator section present on the right side of the figure given above. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(f1) = P(f2) =……=P(fl), and we would conclude that, we maximize P(X/fi). In the other case we maximize P(X/fi)P(fi) that is the overall expression of the numerator.

4. D:Neural Networks.

In concern to biology, the largest cell is "neuron" that performs millions of operations and its components like dendrites, axon etc. plays a really vital role. In terms in artificial neural network the word "perceptron" is a measure of observation.

Processing elements are interconnected with each other and performs a gradient descent methods. These processing components is considered as "neurons". The skilled or the competent neural networks allows the interoperability of the network that has the learning ability. In case of realization of logic gates also, learning is applied to cope up with the manual solution (hit and trial) as it was in McCulloch Pitts Model.

ISSN 2348 – 8034
Impact Factor- 5.070

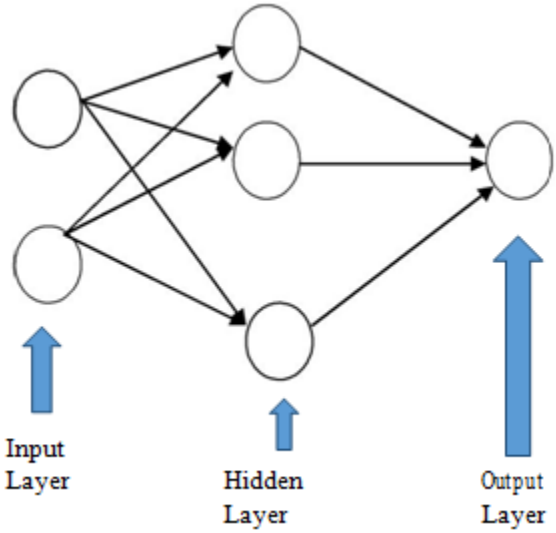

Fig. 5 –Classifier in an approach to NN

For the purpose of pattern recognition and classification neural networks are essentially used. Realization of logic gates can also be done using the neural networks. An example of it was McCulloch Pitts. Minimization of error takes place when weights are adjusted in a nominal fashion.Backpropagation algorithm plays a major role in finding the gradient of the error.In the phase of acquiring new information and problem solving, the processed data that go through the network corresponds to the weight adjustment scenario. In the multiclass scenario of neural network abbreviated as NN, the difficultycanbe marked by making use of the feed forward technique in multi-layer division, where a single neuron is not used. The other types of neuron are as single layer feed forward and feedback networks.

4. E:Support Vector Machine (SVM)

For the case of discrete data i.e. classification, continuous data i.e. regression or general pattern recognition components,Support vector machine plays a vital role. In case of machine learning also, this is used. SVM is examined as anappropriate classifying mechanism because it has an excellent capability of generalization. This happens so because, it does all this beyond the essentiality to count a prior learning with experience considered as knowledge, even when the proportion of the spaceof input seems enormous. When a dataset is considered that is to be linearly separable, a classification function that is linear approaches to a dividing hyper plane f(x) and it goes throughin between two classes, performing the separationand making the one into two. In the beginning when SVM was made it primarily dealt with the binary class problems but slowly and gradually it started working with the multi class problem.

V. PROS & CONS.

Each and every scenario that is ever generated has to be made with some pros and cons. And it is demonstrated in a tabular form below:

ISSN 2348 – 8034
Impact Factor- 5.070

S.No.	Model	Advantage	Disadvantage
1	Decision Trees	Can be combined with other techniques, Simple, understandable, Interpretable, Allowed additional scenarios	Unstable, relatively inaccurate, Somehow biased, Calculations are complex.
2	K-Nearest Neighbor	Implementation is simple, Flexible, Handles multi class cases, Can do well in cases with enough data.	Large search problem to find neighbors that are nearest, storage of data is a peculiar scenario, Distances must be meaningful.
3	Support Vector Machines	Speed and Memory is good.	Choice of the kernel is not appropriate in many cases, Size and speed required in both testing and training, Discrete data is another problem, High Algorithmic complexity.
4	Naive Bayesian Classifiers	Easy to implement, Small amount of training data is required, Good results can be added in most of the cases.	Dependencies exist among variables, Cannot be modelled by Naïve Bayes Classifier.
5	Neural Networks	Ability is present for learning, Can be generalized, Does not impose restrictions on input variables .	Black Box, Extracting knowledge is difficult.

ISSN 2348 – 8034
Impact Factor- 5.070

VI. CONCLUSION

There are many techniques in concern to data mining i.e. in performing data mining many factors plays an important role, and every factor has its own pros and cons. Some of the well-known examples of the algorithms that use the tuples in concern to training to create a generalized model are Bayesian classifiers, Back propagation etc.

Lazy learners are also there such as nearest-neighbor classifiers abbreviated as (NN-classifier) and case-based reasoning has its concern too. In these cases, the space of the pattern contains the tuples and they stall until they are made front with a tuplethat has to be test prior to the abstraction in a conceptual manner.

REFERENCES

1. *Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease prediction system using associative classification and genetic algorithm. arXiv preprint arXiv:1303.5919.*
2. *Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using artificial neural network and feature subset selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence, 13(3).*
3. *Nithya, N. S., &Duraiswamy, K. (2014). Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface. Sadhana, 39(1), 39-52.*
4. *Akinola, S. O., &Oyabugbe, O. J. (2015). Accuracies and training times of data mining classification algorithms: an empirical comparative study. Journal of Software Engineering and Applications, 8(09), 470.*
5. *Majali, J., Niranjan, R., Phatak, V., &Tadakhe, O. (2015). Data mining techniques for diagnosis and prognosis of cancer. International Journal of Advanced Research in Computer and Communication Engineering, 4(3), 613-6.*
6. *Salvithal, N. N., &Kulkarni, R. B. (2016). Appraisal Management System using Data mining Classification Technique. International Journal of Computer Applications (0975–8887) Volume.*
7. *Tanvi Sharma, Anand Sharma &VibhakarMansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.*
8. *Jeetha, B. R. (2014). Efficient classification method for large dataset by assigning the key value in clustering. International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN, 319-324.*
9. *Tomar, D., &Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.*
10. *Krishnaiah, V., Narsimha, D. G., & Chandra, D. N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies, 4(1), 39-45.*