RESEARCH ARTICLE OPEN ACCESS

Hypothesis on Different Data Mining Algorithms

Shraddha Deshmukh*, Swati Shinde**
*(Department of Information Technology, University of Pune, Pune - 05
** (Department of Information Technology, University of Pune, Pune - 05

ABSTRACT

In this paper, different classification algorithms for data mining are discussed. Data Mining is about explaining the past & predicting the future by means of data analysis. Classification is a task of data mining, which categories data based on numerical or categorical variables. To classify the data many algorithms are proposed, out of them five algorithms are comparatively studied for data mining through classification. There are four different classification approaches namely Frequency Table, Covariance Matrix, Similarity Functions & Others. As work for research on classification methods, algorithms like Naive Bayesian, K Nearest Neighbors, Decision Tree, Artificial Neural Network & Support Vector Machine are studied & examined using benchmark datasets like Iris & Lung Cancer.

Keywords - *Artificial Neural Network, Classification, Data Mining, Decision Tree, K-Nearest Neighbors, Naive Bayesian & Support Vector Machine.*

I. INTRODUCTION

Nowadays large amount of data is being gathered and stored in databases everywhere across the globe and it is increasing continuously. Different organizations & research centers are having data in terabytes. That is over 1000 Terabytes of data. So, we need to mine those databases for better use. Data Mining is about explaining the past & predicting the future. Data mining is a collaborative field which combines technologies like statistics, machine learning, artificial intelligence & database. The importance of data mining applications is predicted to be huge. Many organizations have collected tremendous data over years of operation & data mining is process of knowledge extraction from gathered data. The organizations are then able to use the extracted knowledge for more clients, sales & greater profits. This is also true in the engineering & medical fields.

1.1 DATA MINING

Data mining is process of organising available data in useful format. Fig.1 shows basic concept of data mining. Basic terms in data mining are:

- *Statistics*: The science of collecting, classifying, summarizing, organizing, analysing & interpreting data.
- *Artificial Intelligence*: The study of computer algorithms which simulates intelligent behaviours for execution of special activities.
- *Machine Learning:* The study of computer algorithms to grasp the experiences and use it for computerization.
- *Database*: The science & technology of collecting, storing & managing data so users can retrieve, insert, modify or delete such data.
- *Data warehousing*: The science & technology of collecting, storing & managing data with advanced multi-dimensional reporting services in support of the decision making processes.
- *Predicting The Future*: Data mining predicts the future by means of modelling.
- *Modelling*: Modelling is the process in which classification model is created to predict an outcome.

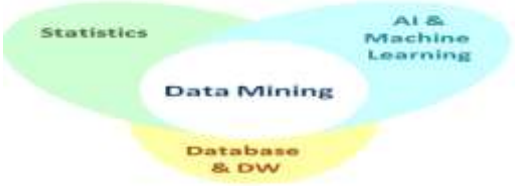


Fig. 1. Concept of data mining

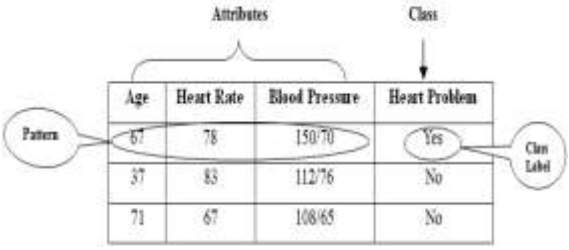
II. CLASSIFICATION

Classification is a data mining task of predicting the value of a categorical variable (target or class) by building a model based on one or more algebraic and/or categorical variables (predictors or attributes). It Classifies data based on the training set & class labels. Examples:

- Classifying patients by their symptoms,
- Classifying goods by their properties, etc.

There are some common terms used in classification process. Table 1 illustrates basic terms used in classification process like pattern (records, rows), attributes (dimensions, columns), class (output column) and class label (tag of class):

TABLE -1: Terms Used in Classification


Attributes
Class
Pattern
Class Label

Age	Heart Rate	Blood Pressure	Heart Problem
67	78	150/70	Yes
37	83	112/76	No
71	67	108/65	No

Classification is a method of data mining for predicting the value for data instances by using previous experiences. Since we want to predict either a positive or a negative response, we will build a binary classification model. Classification is important because it helps scientists to clearly diagnose problems, study & observe them & organize concentrated conservation efforts. It also assists as a way of remembering & differentiating the types of symptoms, making predictions about diseases of the same type, classifying the relationship between different defects & providing precise names for diseases.

2.1 Applications of Classification

Classification have several applications like Medical Diagnosis, Breast Cancer Diagnosis, Market Targeting, Image processing, Wine Classification, Solid Classification for selection of fertilizer, etc.

III. CLASSIFICATION ALGORITHMS

There is quite a lot of research on algorithms that classifies data. Several approaches have been developed for classification in data mining. Fig 2 shows hierarchy of classification algorithms:

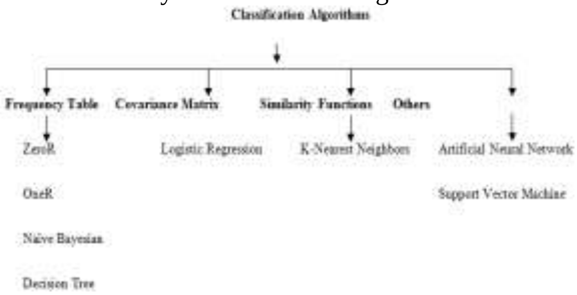


Fig. 2. Hierarchy of classification algorithms

IV. NAÏVE BAYESIAN

4.1 Introduction to Naïve Bayesian

The Naive Bayesian (NB) method is a simple probabilistic classifier based on Bayes Theorem (from Bayesian statistics) with strong (naive) independence premises which assumes that all the features are unique. NB model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets [1].

In the NB classifiers, every feature can help determining which topic should be appointed to a given input value. To choose a topic for an input value, the naive Bayes classifier begins by evaluating the prior probability of each topic, which is determined by checking the frequency of each topic in the training set. The input from each feature is then mixed with this previous probability, to arrive at a probability estimate for each topic. If the estimated probability is the highest is then assigned to the testing inputs [5].

A supervised classifier is built on training corpora containing the correct topic for each input. The framework used by Bayesian classification is shown in Fig.3

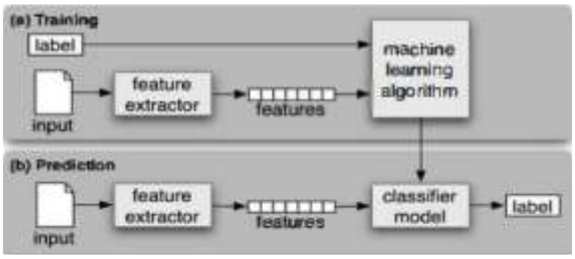


Fig. 3. Bayesian classification

(a) During the training, a feature extractor is used to convert each input value to a feature set. These feature sets, which capture the basic information about each input that should be used to classify it, are discussed in the next section. Pairs of feature sets & topics are fed into the machine learning algorithm to generate a model. (b) During the prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted topics [5].

4.2 Advantages to Naive Bayesian

1. Fast to train & classify
2. Not sensitive to irrelevant features
3. Handles real & discrete data
4. Handles streaming data well

4.3 Disadvantages to Naive Bayesian

1. Assumes autonomy of aspects
2. Dependencies exist among variables (ex. Hospital: Patients, Diseases: Diabetes, Cancer) are not modelled by NB.

V. K NEAREST NEIGHBOR

5.1 Introduction to K-Nearest Neighbor

K nearest neighbor (KNN) is a simple method that stores all available cases & classifies new cases based on a similarity measure (e.g. euclidean). KNN has been used in statistical estimation & pattern recognition already in the beginning of 1970's as a non-parametric technique. The closest neighbor rule distinguishes the classification of unknown data point on the basis of its closest neighbor whose class is already known [4].

The training points are selected weights according to their distances from sample data point. But at the same time the computational complexity & memory requirements remain the essential things. To overcome from memory restriction size of dataset is minimized. For this the repeated patterns which don't include additional data are also excluded from training data set. To further strengthen the information focuses which don't influence the result are additionally eliminated from training data set [4]. The NN training dataset can be formed for utilizing different systems to boost over memory restriction of KNN. The KNN implementation can be done using ball tree, k-d tree, nearest feature line (NFL), principal axis search tree & orthogonal search tree.

Next, the tree structured training data is divided into nodes & techniques like NFL & tunable metric divide the training data set according to planes. The speed of basic KNN algorithm can be increase by using these algorithms. Consider that an object is sampled with a set of different attributes.

5.2 Advantages to K-Nearest Neighbors

1. No assumptions about the characteristics of the concepts to learn have to be done
2. Complex concepts can be learned by local approximation using simple procedures
3. Very simple classifier that works well on basic recognition problems.

5.3 Disadvantages to K-Nearest Neighbors

1. The model cannot be interpreted
2. It is computationally expensive to find the KNN when the dataset is very large
3. It is a lazy learner; i.e. it does not learn anything from the training data & simply uses the training data itself for classification.

VI. Decision Tree

6.1 Introduction to Decision Tree

A decision tree (DT) is a decision support tool that uses a tree-like graph or model of decisions & their possible effects, including chance event results, assets cost & utility. It is the only way to display an algorithm.

The decision tree is a method for information portrayal evolved in the 60s. It can resolve the class label of test patterns by using set value of attribute. DT is a cycle free graph which has nodes as attributes to support decisions. The tree branch represents a precedence connection between the nodes [6]. The value of a branch is an element of the attribute value set of the branch's parent node. The attributes are nodes with at least two children, because an attribute has got as many branches as the cardinality of the value set of the actual attribute. The root of the tree is the common ancestor attribute, from where the classification can be started. The leaves represent class nodes of the tree. In every relation the class is only a child, so it is a leaf of the tree in every case [6].

DT builds classification model in the form of a tree. It breaks a dataset into smaller subsets and an associated decision tree is incrementally developed. The final result is a tree with decision nodes & leaf nodes. A decision node has two or more branches and leaf node represents a decision. The topmost node also called as root node which corresponds to the best predictor. DT can handle both categorical & numerical data [4]. DT helps formalize the brainstorming process so we can identify more potential solutions.

6.2 Advantages to Decision Tree

1. Easy to interpretation.
2. Help determine worst, best & expected values for different scenarios.

6.3 Disadvantages to Decision Tree

1. Determination can get very complex if many values are ambiguous.

VII. Artificial Neural Network

6.1 Introduction to Artificial Neural Network

Artificial neural networks (ANNs) are types of computer architecture inspired by nervous systems of the brain & are used to approximate functions that can depend on a large number of inputs & are generally unknown. ANN are presented as systems of interconnected "neurons" which can compute values from inputs & are capable of machine learning as well as pattern recognition due their adaptive nature [4]. The brain basically learns from experience. It is natural proof that some problems that are beyond the scope of current computers are indeed solvable by small energy efficient packages. This brain modeling also promises a less technical way to develop machine solutions. This new approach to computing also provides a more graceful degradation during system overload than its more traditional counterparts [8].

A neural network is a massively parallel-distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge & making it available for use. Neural network are also referred to in literature as neuro computers, connectionist networks, parallel-distributed processors, etc. A typical neural network is shown in the fig. 4.Where input, hidden & output layers are arranged in a feed forward manner [8]. The neurons are strongly interconnected & organized into different layers. The input layer receives the input & the output layer produces the final output. In general one or more hidden layers are sandwiched in between the two [4]. This structure makes it impossible to forecast or know the exact flow of data.

Input Layer Hidden Layer Output Layer

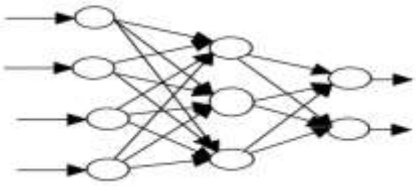


Fig. 4. A simple neural network

ANN typically starts out with randomized weights for all their neurons. This means that initially they must be trained to solve the particular problem for which they are proposed. During the training period, we can evaluate whether the ANN's output is correct by observing pattern. If it's correct the neural weightings that produced that output are reinforced; if the output is incorrect, those weightings responsible can be diminished [4]. An ANN is useful in a variety of real-world applications such as visual pattern recognition, speech recognition and programs for text-to-speech that deal with complex often incomplete data.

6.2 Advantages to Artificial Neural Network

1. It is easy to use, with few parameters to adjust.
2. A neural network learns & reprogramming is not needed.
3. Applicable to a wide range of problems in real life.

6.3 Disadvantages to Artificial Neural Network

1. Requires high processing time if neural network is large.
2. Learning can be slow.

VIII. SUPPORT VECTOR MACHINE

8.1 Introduction to Support Vector Machine

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

The beauty of SVM is that if the data is linearly separable, there is a unique global minimum value. An ideal SVM analysis should produce a hyperplane that completely separates the vectors (cases) into two non-overlapping classes. However, perfect separation may not be possible, or it may result in a model with so many cases that the model does not classify correctly. In this situation SVM finds the hyperplane that maximizes the margin & minimizes the misclassifications [4].The algorithm tries to maintain the slack variable to zero while maximizing margin. However, it does not minimize the number of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes [9].

The simplest way to separate two groups of data is with a straight line (1 dimension), flat plane (2 dimensions) or an N-dimensional hyperplane as shown in Fig. 5.

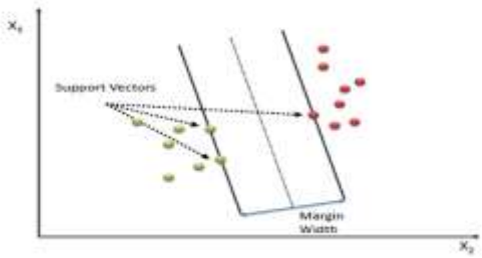


Fig. 5. Hyperplane in SVM

The main reason you would want to use an SVM instead of a Logistic Regression is because problem might not be linearly separable and if you are in a highly dimensional space [9].

8.2 Advantages to Support Vector Machine

1. Most robust & accurate classification technique for binary problems.
2. Memory-intensive.

8.3 Disadvantages to Support Vector Machine

1. It can be painfully inefficient to train
2. High complexity & extensive memory requirements for classification in many cases.
3. Supports only binary classification.

IX. Comparative Study of Algorithms

As per comparative analysis from Table 2 we can say that, ANN is better for data classification. Neural network allows learning from experiences & supports decision making, classification, Pattern recognition, etc. Neural networks often exhibit patterns similar to those exhibited by humans. However this is more of interest in cognitive sciences than for practical examples.

TABLE -2: Comparative Study of Algorithms

Algo.	**Approach**	**Features**	**Flaws**
NB	Frequency Table	• Simple to implement. • Great Computational efficiency & classification rate. • It predicts accurate results for most of the classification & prediction problems.	• The precision of algorithm decreases if the amount of data is less. • For obtaining good results it requires a very large number of records.
KNN	Similarity Function	• Classes need not be linearly separable. • Zero cost of the learning process. • Well suited for multimodal classes.	• Time to find the nearest Neighbors in a large training data set can be excessive. • Performance of algorithm depends on the number of dimensions used
DT	Frequency Table	• It produces the more accuracy result than the C4.5 algorithm. • Detection rate is increase & space consumption is reduced.	• Requires large searching time. • Sometimes it may generate very long rules which are very hard to prune. • Requires large amount of memory to store tree.
ANN	Others	• It is easy to use & implement, with few parameters to adjust. • A neural network learns & reprogramming is not needed. • Applicable to a wide range of problems in real life.	• Requires high processing time if neural network is large. • Learning can be slow.
SVM	Others	• High accuracy. • Work well even if data is not linearly separable in the base feature space.	• Speed & size requirement both in training & testing is more. • High complexity & extensive memory requirements for classification in many cases.

X. Experimental Analysis

To examine all studied methods Lung Cancer & Iris benchmark datasets are used. Results are for 100% training & 100% testing scenario. Result analysis on the basis of accuracy is given in Table 3. Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Total number of correctly classified data}}{\text{Total number of data}}$$

TABLE -3: Result Analysis of Methods Using WEKA Tool

	Iris	Lung Cancer
NB	94.7%	97.4%
KNN	94.0%	97.0%
DT	95.0%	98.0%
SVM	65.0%	98.5%
ANN	98.3%	100%

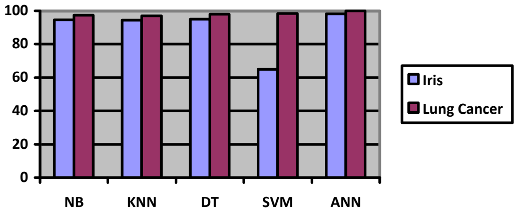


Fig 6: Chart of Result Analysis

XI. CONCLUSION

In this paper, we have studied five different classification methods based on approaches like, frequency tables, covariance matrix, similarity functions & others. Those algorithms are NB, KNN, DT, ANN & SVM. As per comparative study in between all algorithms we reached to conclusion that ANN is most suitable & efficient technique for Classification. ANNs are considered as simplified mathematical models of human brain & they function as parallel distributed computing networks. ANNs are universal function approximates, & usually they deliver good performance in applications. ANN has generalization ability as well as learnability. It is easy to use, implement & applicable to real world problems.

There is a huge scope in this area of classification by using different methods of ANN like Fuzzy with ANN, Neuro-fuzzy, Genetic Approach, etc. in Artificial Neural Network.

REFERENCES

Journal Papers:

[1] M Ozaki, Y. Adachi, Y. Iwahori, and N. Ishii, Application of fuzzy theory to writer recognition of Chinese characters, *International Journal of Modeling and Simulation, 18(2),* 1998, 112-116.

[2] R. Andrews, J. Diederich & A. B., Tickle, "Survey & critique of techniques for extracting rules from trained artificial neural networks," *Knowledge Based System, vol. 8, no. 6,* pp. 373-389, 1995.

[3] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", *Journal of Software Engineering & Applications, 6, 85-97*, 2013.

[4] Wang Xin, Yu Hongliang, Zhang Lin, Huang Chaoming, Duan Jing, "Improved Naive Bayesian Classifier Method & the Application in Diesel Engine Valve Fault Diagnostic", *Third International Conference On Measuring Technology & Mechatronics Automation*, 2011.

[5] Sagar S. Nikam, "A Comparative Study Of Classification Techniques In Data Mining Algorithms", *Oriental Journal Of Computer Science & Technology*, April 2015.

[6] Zolboo Damiran, Khuder Altangerelt, "Text Classification Experiments On Mongolian Language", *IEEE Conference*, Jul 2013.

[7] Zolboo Damiran, Khuder Altangerel, "Author Identification: An Experiment based on Mongolian Literature using Decision Tree", *IEEE Conference*, 2013.

[8] Essaid el Haji, Abdellah Azmani, Mohm el Harzli, "A pairing individual-trades system , using KNN method", *IEEE Conference*, 2014.

[9] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, Sarthak K., "A Rainfall Prediction Model using Artificial Neural Network", *IEEE Conference*, 2012.

[10] Erlin, Unang Rio, Rahmiati, "Text Message Categorization of Collaborative Learning Skills in Online Discussion Using SVM", *IEEE Conference*, 2013.