Survey on Data Mining Classification Algorithms

Mehernaaz Patel[1] Nisha Toke[2] Mugdha Umarjikar[3] Sanket Samaiya[4]
[1,2,3,4]DIT, Pimpri-18, India

Abstract— Data mining classification algorithms are used to find out in which group each data instance is related within a given dataset. These algorithms used for classifying data into different classes according to some constrains. Several major kinds of classification algorithms including C4.5 [2], ID3 and Naive Bayes are used for classification [1][3]. Factors that affect the performance of a classification algorithm are training data set, number of tuples and attributes, types of attributes and system configuration. While considering these factors this paper provides an inclusive study of different classification algorithms and their features and limitations.

Key words: Data Mining, Classification, ID3, Naive Bayes, C4.5

I. INTRODUCTION

Data Mining is the process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases and are hidden among large amounts of data.

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedure is recognized method for repeatedly making such decisions in new situations. Here if we assume that problem is a concern with the construction of a procedure that will be applied to a continuing sequence of cases in which each new case must be assigned to one of a set of pre-defined classes on the basis of observed features of data. Creation of a classification procedure from a set of data for which the exact classes are known in advance is termed as pattern recognition or supervised learning. Contexts in which a classification task is fundamental include, for example, assigning individuals to credit status on the basis of financial and other personal information, and the initial diagnosis of a patient's disease in order to select immediate treatment while awaiting perfect test results. Some of the most critical problems arising in science, industry and commerce can be called as classification or decision problems. Three main historical strands of research can be identified: statistical, machine learning and neural network. All groups have some objectives in common. They have all attempted to develop procedures that would be able to handle a wide variety of problems and to be extremely general used in practical settings with proven success.

This paper gives a comparative study on classification algorithms in data mining. Following algorithms are studied:

- ID3 Algorithm
- C4.5 Algorithm
- Naïve Bayes Algorithm

II. METHODOLOGY

A. ID3

ID3 decision tree algorithm is a classic algorithm, it starts from the root node. Root node is one of the best attributes. Then the property values are generated corresponding to each branch. Each branch has generated new node. For the best attributes of the selection criteria, ID3 using entropy-based definition of information gain to select the test attribute within the node. Entropy characterizes the purity of any sample set.

Suppose S is a set s of data samples. Assume that class label attribute has m different values, definition of m different classes Ci (i=1…m). Set Si is the number of samples in class Ci. Equation (1) is on a given sample classification to the expectations of the information.

$$I(s_1,s_2,...,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

Where Pi is the probability of any sample belonging to Ci, with si /s estimated.

Set attribute A has v different values {a1, a2,… , av}.A property can be divided into v subsets S{S1,S2,…,Sv} Where, Sj contains a number of S in this sample, They have a value of aj in A. If selected as test attribute A, These subsets correspond to the set S contains nodes from the growth of branching out. S_j assumption S_{ij} is a subset of the samples of class C_i. According to the A divided into subsets of entropy or expected information is given by Equation (2):

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j}+\cdots+s_{mj}}{s} I(s_{1j},...,s_{mj}). \quad (2)$$

Item $(s_{1j}+...+s_{mj})/s$ sub-set as the right of first j, and is equal to the number of subset of the sample divided by the total number of S in the sample. Equation (3) is a given subset of S_j.

$$I(s_{1j},s_{2j},...,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3)$$

Where $p_{ij}=s_{ij} / |s_j|$ is a sample of S_j in the probability of belonging to class C_i. Equation (4) is a branch will be in the encoding information.

$$Gain(A) = I(s_1,s_2,...s_m) - E(A). \quad (4)$$

In other words, Gain (A) is due to the value of that property a result of the expectations of the entropy of compression.

From this, the smaller the entropy value, the lower the correlation, a subset of the division of the higher purity,

the higher the corresponding information gain. Therefore, the test attribute decision tree selected for the properties with the highest information gain. It creates a node and to mark the property, each value of the property to create branch and accordingly divided the sample.

B. C4.5

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector, where they represent attribute values or features of the sample, as well as the class in which falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The information entropy is calculated according to equation 1.

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \frac{|\{x \in Ex \mid value(x,a) = v\}|}{|Ex|} H(\{x \in Ex \mid value(x,a) = v\}) \quad (1)$$

Information entropy calculation formula is shown in equation (2):

$$\inf o(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

Calculation formula of information gain

$$IGR(Ex,a) = IG / IV \quad (3)$$

rate is:

Where, IG is the information gain in front. And IV calculation formula is as follows

$$IG(Ex,a) = \sum_{v \in values(a)} \frac{|\{x \in Ex \mid value(x,a) = v\}|}{|Ex|} \cdot \log_2\left(\frac{|\{x \in Ex \mid value(x,a) = v\}|}{|Ex|}\right) \quad (4)$$

The simplified formula is as follows:

$$H(v) = -\sum_{j} p(v_j) \log p(v_j) \quad (5)$$

Where, V represents the full value of an attribute in the attribute set A.

C. Naive Bayes

The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

Why preferred Naive Bayes implementation:

1) When the data is high.
2) When the attributes are independent of each other.
3) When we want more efficient output, as compared to other methods output.

1) Bayes Rule

A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, E, where a dependence relationship exists between C and E.

This probability is denoted as P (C |E) where

$$P(C \mid E) = \frac{P(E \mid C)P(C)}{P(E)}$$

2) Naive Bayesian Classification Algorithm

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1) Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X= (x_1, x_2…, x_n), depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2..., A_n.
2) Suppose that there are m classes, C_1, C_2…, C_m. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class C_i if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq I$ Thus, we maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

3) As P(X) is constant for all classes, only P ($X|C_i$) P (C_i) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1)=P(C_2)=\ldots=P(C_m)$, and we would therefore maximize P(X|Ci). Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i)=|C_i,D|/|D|$, where $|C_i,D|$ is the number of training tuples of class C_i in D.
4) Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X \mid Ci) = \prod_{k=1}^{n} P(xk \mid Ci)$$

$$=P(x1|C_i)x\ P(x2|C_i)x\ldots\ P(x_m|C_i).$$

We can easily estimate the probabilities P(x1|Ci), P(x2|Ci),… ,P(x_m|C_i) from the training tuples. Recall that here Xu refers to the value of attribute AK for tuple X. For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute P(X|Ci), we consider the following:

a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci,D|, the number of tuples of class Ci in D.
b) If Ak is continuous valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

So that

$$P(xk|C_i)=g(xk,\ \mu ci,\ \sigma ci)$$

We need to compute μci and σci, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci. We then plug these two quantities into the above equation.

5) In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

$P(X|Ci)P(Ci)>P(X|Cj)P(Cj)$ for $1 \le j \le m, j \ne i$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.

III. Literature Survey

Wang Xiaohu1, Wang Lele2, Li Nianfeng2 [1], this paper deals with the application of classical decision tree ID3 of the data mining in a certain site data. It constitutes a decision tree based on information gain and thus produces some useful purchasing behaviour rules. ID3 is a classic algorithm that starts from root node. In this entropy characterizes the purity of any sample set. ID3 algorithm cannot handle noisy data. Here, missing data cannot be mined.

Xuefei Wang1, Yan Shi2 [2], this paper proposed a new method for targeting advertising based on C4.5 algorithm and cloud storage. The mess data about users, such as user information, user browsing behaviour and so on, is stored in the cloud. With C4.5 algorithm, we do mess data processing for data classification. C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao [3], in this paper a Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

Sunita Joshi* Bhuwaneshwari Pandey Nitin Joshi [4], in this study we focused on comparison of two classification techniques and few issues like accuracy and cost. Time to build the model is less when using Naive Bayes and correctly classified instances are more when using Naive Bayes and prediction accuracy is also greater in Naive Bayes than of J48 (extension of c4.5). Hence Naive Bayes is a better algorithm than J48.

Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI [5], in this paper, focus is on the key elements of their construction from a set of data, then the author presented the algorithm ID3 and C4.5 that respond to these specifications. And comparison of ID3/C4.5, C4.5/C5.0 and C5.0/CART is being shown, which led us to confirm that the most powerful and preferred method in machine learning is certainly C4.5. ID3 is overly sensitive to feature with large number of values. It builds decision tree from fixed set of values. C4.5 prunes tree using single pass algorithm. Both are less accurate.

IV. Conclusion

By performing this study on data mining classification algorithms, we came to know the method, advantages and disadvantages of ID3, C4.5 and Naive Bayes Classification algorithms. This study helps in choosing the correct algorithm for various data mining applications.

References

[1] Wang Xiaohu1, Wang Lele2, Li Nianfeng2: An Application of Decision Tree Based on ID3. 2012 International Conference on Solid State Devices and Materials Science.
[2] Xuefei Wang1, Yan Shi2: Design and Implementation of Targeting Advertising System based on C4.5 Algorithm. 2015 4th International Conference on Computer Science and Network Technology (ICCSNT 2015).
[3] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao: Decision Support in Heart Disease Prediction System using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE)
[4] Sunita Joshi* Bhuwaneshwari Pandey Nitin Joshi: Comparative analysis of Naive Bayes and J48 Classification Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering.
[5] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI: A comparative study of decision tree ID3 and C4.5. (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.

