RESEARCH ARTICLE OPEN ACCESS

Pre-processing data using ID3 classifier

Hemangi Bhalekar[1], Swati Kumbhar[2], Hiral Mewada[3], Pratibha Pokharkar[4], Shriya Patil[5],Mrs. Renuka Gound[6].
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

Abstract:

A Decision tree is termed as good DT when it has small size and when new data is introduced it can be classified accurately. Pre-processing the input data is one of the good approaches for generating a good DT. When different data pre-processing methods are used with the combination of DT classifier it evaluates to give high performance. This paper involves the accuracy variation in the ID3 classifier when used in combination with different data pre-processing and feature selection method. The performances of DTs are produced from comparison of original and pre-processed input data and experimental results are shown by using standard decision tree algorithm-ID3 on a dataset.

Keywords — Decision Tree, ID3, Feature selection

I. INTRODUCTION

Decision tree gives alternative option between two branch nodes, and its leaf node represents a decision. Decision making becomes easier by human when a DT is small in size. But it is not possible in case of big data, as there are numbers of attributes present which can have noisy, incomplete or inadequate data, which cannot be classified easily. The second constraint is an accuracy of the DT, accuracy can be measured when new data is introduced to classifier, and it predicts accurate class for sample data by using the defined DT rules. It is not always possible to get small sized tree which predicts accurate class for new data. In general there can be two approaches for this, either generates a new decision tree algorithm that can predict the data accurately or perform pre-processing methods on input data such as normalization, numeric to binary, discretize. In this paper, we are focusing on the second approach that is pre-processing the training data. The following process describes the flow.

This paper describes, how pre-processing is beneficial when used with feature selection methods and ID3 classifier. The data is selected from database for performing training the data. On that data various pre-processing methods are applied to get the data with high accuracy. This data is then passed to an ID3 classifier for processing. Classifier then processes it, and generates the output. It decides which attribute should be root node and which should be leaf and it is calculated using the formulas like entropy and information gain.

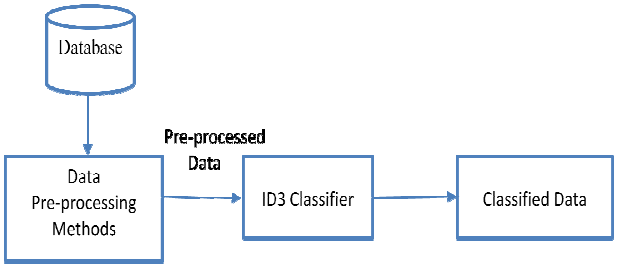

Fig 1: Flow of classification

II. LITERATURE SURVEY

Data pre-processing is one of the methods used for the data handling in big data era. There have been many concepts of pre- processing that are been proposed for the proper understanding of the concepts and techniques. Considering the data mining and pre processing are the most necessary parts to be focused. The various techniques or the steps that are included in the pre-processing are explained in a proper manner in [1] where along with it the data mining concepts are also elaborated in a proper method. The machine learning algorithms may include Decision Tree, neural networks, nearest neighbour algorithm, etc which can be referenced by [2] this gives an elaborated view of building the decision tree. For generating a decision tree with accurate measures, the best approach can be considered as the training data pre-processing. For this purpose there is the need of attribute selection method and attribute generation. Formulae are to be used and algorithms to be implemented for such attribute selection and generation methods which are completely demonstrated in [3]. Empirical experiments are performed for the confirmation of the effectiveness of the proposed theory. Here the threshold value is considered for the association rule generation.

The two entropy-based methods i.e. C4.5 and ID3 for which the pre-processed data and the original data are induced from the decision trees. The methods that are explained [4] may tell us that the collections of attributes that describe objects are extended by the new attributes and secondly, the original attributes are replaced by the new attributes. For machine learning pre processing can also be conducted based on the formal concept analysis, both of the above mentioned methods consider the Boolean factor analysis determined in [4]. Data preparations and various data filtering steps require a lot of time in machine learning. By the help of data pre processing the noisy, unwanted, irrelevant data can be removed which makes it easier for further work. The pre-processing of data may include data cleaning, normalization, transformation, feature extraction and selection, etc. Once the data gets pre-processed, the outcome is treated as a final training set. It is not applicable that only single sequence of algorithms for data pre-processing gives a best performance and hence various data pre-processing algorithms can be known [5]. There the paper provides complete sectional explanation for data pre-processing by using various methodologies and algorithms. The supervised learning algorithms include Naïve Bayes, C4.5 decision tree algorithm. Ranking algorithms help to reduce the dimensionality of the feature selection and classification.[6]. The Weka tool can be proved useful for data pre-processing, classification, clustering.[7]

III. FEATURE SELECTION

Feature Selection is a term in data mining that generally used to describe the techniques and methods to reduce the input to a manageable size for processing. Feature selection selects a subset which provides maximum information about a dataset, and removes all variable which gives minimum or unpredictable information of a dataset. Hence, it is also termed as attribute subset selection or variable selection method. It is difficult to find correct predictive subset, so it itself becomes a problem. For e.g. an expert programmer can tell just by seeing the code, that what output it may generate and what possible errors will be there after compilation. It has three main advantages improved model interpretability, shorter training times, enhanced generalization by reducing over fitting.
Algorithms used for selection process can be categorized as wrappers, filters, and embedded methods. Filter method attempt to assess the merits of attributes from the data, ignoring learning algorithm. E.g. Information Gain. Wrapper methods the attributes subset selection is done using the learning algorithm as a black box. E.g. recursive feature elimination algorithm. And embedded method learns while creating a model, which feature gives best accuracy of the model. E.g. Elastic Net, Ridge Regression. In this paper we evaluate Id3 classifier with following feature selection.

A. Correlation feature selection (CFS):

This method describes a subset containing attributes that are highly correlated to class but uncorrelated with each other.

Heuristic merit for subset S formalized as:

$$Merit_s = \frac{k\overline{r_{ca}}}{\sqrt{k + k(k-1)\overline{r_{aa}}}}$$

Where k: number of attributes.

r_{ca} : Average attribute-class correlation

r_{aa}: Average attribute-attribute correlation

B. Entropy

It is introduced by Claude Shannon in 1948. Entropy is a measure of how certain or uncertain the value of a random variable is. Lower value implies less uncertainty, while higher value implies more uncertainty.

$$H(X) = -\sum_{i=1}^{n} p_i log_2 p_i$$

Where, p_i is the probability that an arbitrary tuple X belongs to class.

C. Information Gain

Information Gain measures the change in entropy after making a decision on the value of an attribute. For decision trees, it's ideal to base decisions on the attribute that provides the largest change in entropy, the attribute with the highest gain.

In decision tree, decisions are based on the largest change in entropy, with the attribute which has highest information gain. This becomes a root node initially, and then further decision nodes are calculated by the same approach. The information gain, Gain(S,A) of an attribute A, relative to a collection of examples S, is defined as

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

i. Best First Feature Selection:

This method searches the space of the attribute subsets with backtracking. It may start with empty set of attributes and will start searching in forward for the accurate feature. If there is full set of attributes, searching may start in backward. Or if it starts in between the attribute set, then it searches in both forward and backward direction.

ii. Greedy Stepwise Feature Selection:

Greedy stepwise may start when there are no or all attribute subsets present or it starts in between the attribute subset searching for the best feature. If drop off is there in evaluation of addition/deletion of any remaining attributes greedy terminates it. By traversing the space from one end to other it can produce a ranked list of attributes and records the attribute orders which are selected.

iii. Ranker Feature Selection:

Ranking feature selection does not eliminate any feature, and this filter simply ranks the feature. Based on the relevant information, it prioritizes features and based on this, it allows one to choose feature liable on specified criteria.

IV. DATA PRE-PROCESSING

Pre-processing is an important step in classification algorithms in data mining. Now why to pre-process the data? Data in factual world is dirty, i.e. incomplete means there can be lacking attributes, missing attributes of interest. Data may have noise that may contain error, or inconsistent data which may have discrepancies in code or in names. Even data warehouse requires consistent integration quality data.

Data may have quality problems that need to be addressed before applying a data mining technique. Pre-processing may be needed to make data more suitable for data mining. If you want to find gold dust move the rocks away first, such a way that to get accurate results pre-processing is efficient technique in classification. The process of pre-processing can be categorized as:

1. Data Cleaning: It involves filling of missing values, smoothen the noisy data also involves identifying or removal of outliers

2. Data Integration: This involves integration of multiple databases, data cubes or data files.

3. Data Transformation: This involves normalization and aggregation.

4. Data Reduction: This involves obtaining of reduced representation in volume but it also produces the same or similar analytical results.

This paper studies the effects of following Data pre-processing methods on performance of ID3 classifier.

a. Normalize: This method normalizes all the attributes present in a dataset, and if specified it excludes class attribute.

b. Numeric to Binary: This method converts all the attribute values into binary. If the Numeric value of an attribute is other than 0, then it is converted into binary i.e. 1.

c. Discretize: It reduces the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

d. PKI Discretize: It uses Equal Frequency binning to divide the numeric values, and utilize number of bins which are equal to square root of non-missing values.

e. Standardize: When this pre-processing method is used, the resulted output has zero mean and unit variance.

V. ID3 CLASSIFIER

It stands for Iterative Dichotomizer3, which was invented by Ross Quinlan in 1979. It works by applying the Shannon Entropy, for generating decision tree. It calculates entropy and information gain of each feature and one having highest information gain becomes root node of a tree. Following are the steps for implementing an ID3 classifier

1. Establish classification Attribute(in Table R)
2. Compute classification Entropy.
3. For each attribute in R, calculate Information Gain(discussed above) using classification attribute.
4. Select Attribute with the highest gain to be the next node in the tree(Starting from root node).
5. Remove node attribute, creating reduced table R_s.
6. Repeat steps iii-v until all attributes have been used, or the same classification value remains for all rows in the reduced table.

VI. EXPERIMENTAL RESULTS

In this paper, records belonging to known diabetes dataset were extracted to create training and testing dataset for experimentation. Performance of ID3 classifier in combination with feature selection and data pre-processing methods is evaluated using Diabetes dataset. Following are the results which show accuracy of ID3 classifier when training and testing dataset was pre-processed using various data pre-processing methods.

Fig2 shows that ID3 when combined with normalize pre-processing method provide the accuracy of 92.58%. N2B abbreviation stands for Numeric to Binary in these graphs.

Fig3 shows accuracy of ID3 when a selected feature of training and testing dataset by InfoGain with best first were pre-processed using various data pre-processing methods. These feature selection with Discretize pre-processing method gives the maximum accuracy of 99.59%.

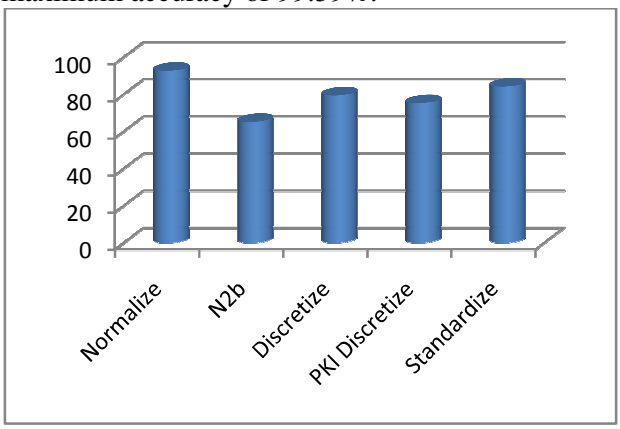

Fig 2: Without Attribute Selection

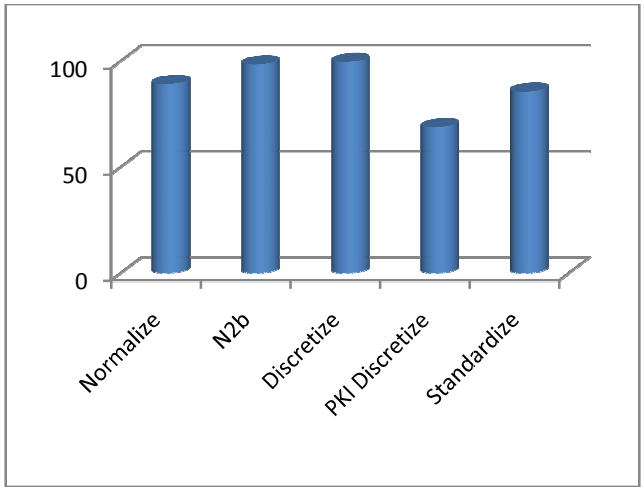

Fig 3: InfoGain with Best First

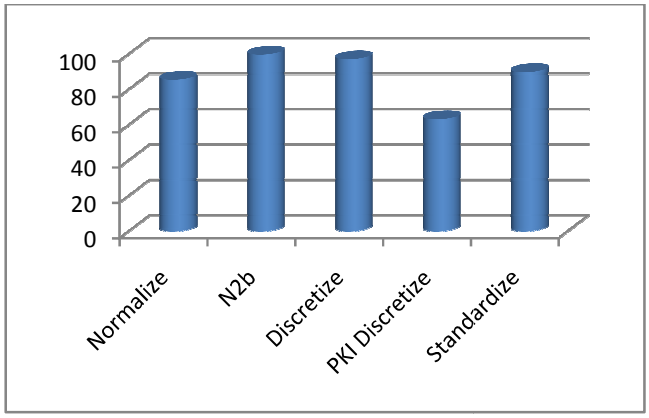

Fig 4: InfoGain with Greedy Stepwise

Fig4 shows maximum accuracy of 99.63% when used with ID3 and selected Greedy Stepwise feature and Numeric to Binary pre-processing method.

Fig5 shows accuracy of ID3 when a selected feature of training and testing dataset by InfoGain with Ranker were pre-processed using various data pre-processing methods. These feature selection in combination with N2B pre-processing method gives the maximum accuracy of 97.76%.

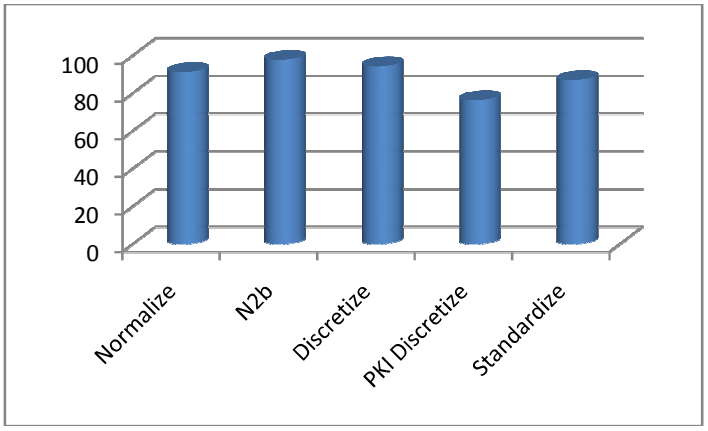

Fig 5: InfoGain with Ranker

VII. COMPARISON

It is observed that when attributes are processed directly by ID3 classifier by without attribute selection, result obtained is 65.1% with N2b and while with Discretize it gives 79.29% accuracy.
When comparison is carried between with and without pre-processing methods it is seen that the accuracy is highly improvised in case of N2B and Discretize pre-processing methods. When N2b is applied with best first it increased up to 98.45% while with ranker it reached up to 97.76% and with Greedy it is up to 99.63%. When Data is pre-processed using best first feature selection with discretize it is increased to 99.29%. When it is combined with Ranker and Discretize it gives 94.3% while with Greedy stepwise in combination with Discretize it gives 96.92% of accuracy.

VIII. CONCLUSION

ID3 classifier performed significantly better when combined with Numeric to Binary data pre-processing method. One of the approaches, for better accuracy, improved ID3 can be used or different set of multi classifiers can be used. But it is observed that, instead of using these approaches one can simply perform data pre-processing

methods on dataset and can achieve better performance in combination with ID3 classifier.

IX. REFERENCES

1) S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas,"Data Preprocessing for Supervised Leaning ",INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME 1 NUMBER 2 2006 ISSN 1306-4428
2) SwastiSinghal, Monika Jena **,"** A Study on WEKA Tool for Data Preprocessing, Classification and Clustering",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013
3) Jasmina NOVAKOVIĆ, Perica STRBAC, Dusan BULATOVIĆ ," TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS, Yugoslav Journal of Operations Research21 (2011), Number 1, 119-135
DOI: 10.2298/YJOR1101119N
4) RanWanga,Yu-LinHeb,∗,Chi-YinChowa,Fang-FangOub,JianZhang,"LearningELMTreefrombig databasedonuncertaintyreduction".
5) Masahiro Terabe,Osamukatai,Tetsuosawaragi,takashiwashio,hiroshimotoda," The data preprocessing method for decision tree using associations in the attributes".
6) Jan Outrata," Preprocessing input data for machine learning by FCA", Department of Computer Science, Palacky University, Olomouc, Czech Republic
T˘r. 17.listopadu 12, 771 46 Olomouc, Czech Republic.
7) KDD,Kdd cup 1999 dataset, <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>, 1999 (accessed January 2013).
8) Hu Zhengbing, Li Zhitang,WuJunqi, "Novel Network IntrusionDetection System(NIDS) Based on Signatures Search of Data Mining",Knowledge Discovery and Data Mining,WKDD 2008. First International Workshop, 2008.
9) Ashish Kumar, Ajay K Sharma, Arun Singh, "Performance Evaluation of BST Multicasting Network over ICMP Ping Flood for DDoS", International Journal of Computer Science & Engineering Technology (IJCSET)
10) P. Arun Raj Kumar , S. Selvakumar, "Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems", Computer Communications, vol. 36, 2013, pp. 303–319.
11) Kejie Lu, Dapeng Wu, Jieyan Fan, SinisaTodorovic, Antonio Nucci, "Robust and efficient detection of DDoS attacks for large-scale internet", Computer Networks, vol. 51, 2007, pp. 5036-5056.