RESEARCH ARTICLE OPEN ACCESS

Selection of Significant Features Using Decision Tree Classifiers

Preeti Kumari[1], K. Rajeswari[2],
[1]M.E Student, PCCOE, Pune
[2]Ph.D Research Scholar, SASTRA and Associate professor, PCCOE, Pune, India

Abstract

Data Mining refers to extraction or mining knowledge from huge volume of data. Classification is an important data mining technique with various applications. It classifies data of various kinds and used to classify each item in a set of data into one of predefined set of classes or groups. This work has been carried out to select the most significant attributes from a dataset using Decision tree classifier j48.This Decision tree classification algorithm can be efficiently used in selecting the most significant feature of a dataset and hence we can reduce the dimensionality of a dataset and yet obtain a very good accuracy. In this paper we have selected two datasets from university of California, Irvine website, of weather and vote and applied j48 algorithm on it. After each iteration, one attribute is removed and its accuracy has been found .We have compared the accuracy of dataset with different attribute count and then selected the most significant features of dataset.

I. Literature review

1.1 Data mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining is a tool for analyzing data allowing users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model to classify a given instance.

1.2 Feature selection

Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. The ability to apply feature selection is critical for effective analysis, because datasets frequently contain far more information than is needed to build the model. For example, a dataset might contain 500 columns that describe the characteristics of customers, but if the data in some of the columns is very sparse you would gain very little benefit from adding them to the model. If you keep the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model. Selection of significant features of a dataset is an important task in any application**.**

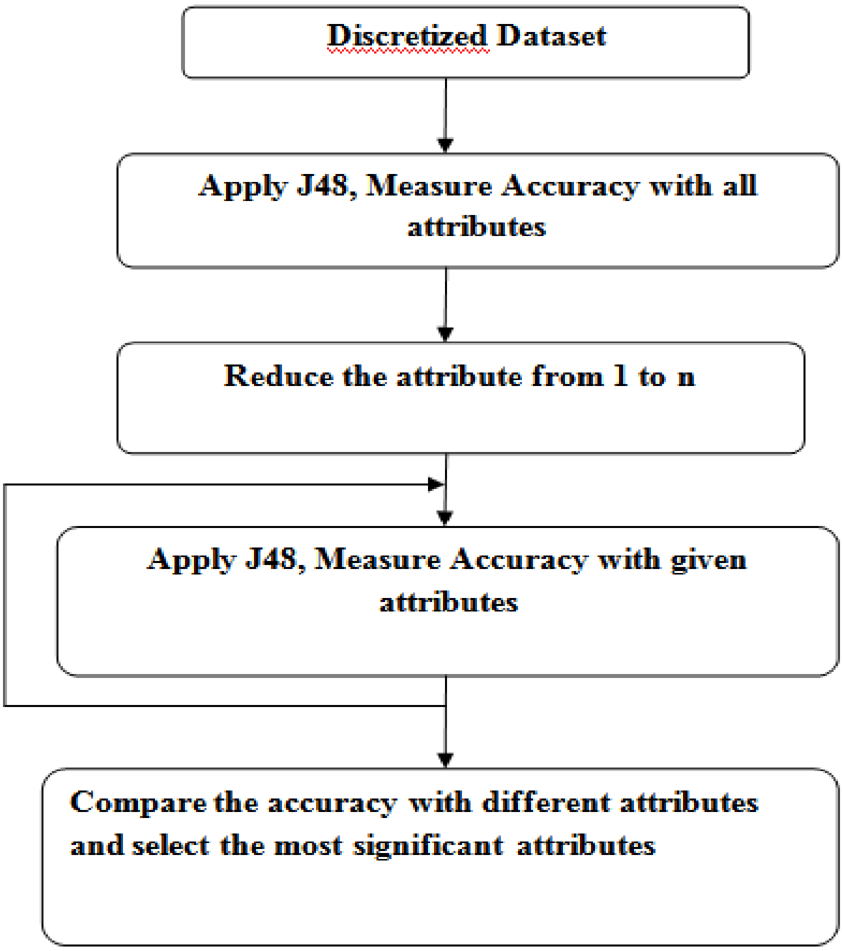
1.3 Decision tree classifier J48

J48 are the improved versions of C4.5 algorithms or can be called as optimized implementation of the C4.5. The output of J48 is the Decision tree. A Decision tree is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree consist a decision and that decision leads to our result. Decision tree divide the input space of a data set into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node.

II. Methodology

In Weka datasets should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file, the Preprocess panel has facilities for importing data from a database, and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria Datasets

lassify tab is used for the classification purpose. We have selected the decision tree classifier J48 to obtain the results.
Step 1: Take the input dataset descretize it properly.
Step 2: Apply the decision tree classifier algorithm J48 on the whole data set and note the accuracy given by it.
Step 3: Reduce the dimension of the given dataset by eliminating one or more attribute pair and note the accuracy of dataset after applying J48 on it.
Step 4: Repeat step 4 for all combination of attributes.
Step 5: Compare the different accuracy provided by the dataset of different attribute taken together and identify the significant feature of the dataset


Discretized Dataset

Apply J48, Measure Accuracy with all attributes

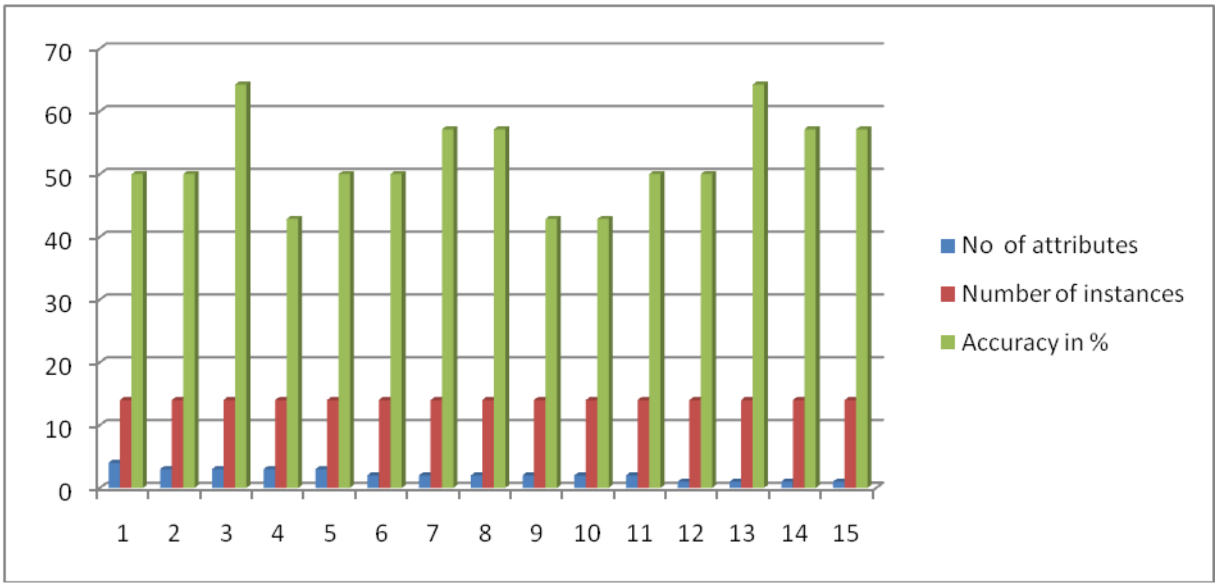
Reduce the attribute from 1 to n

Apply J48, Measure Accuracy with given attributes

Compare the accuracy with different attributes and select the most significant attributes

III. Results and Discussion

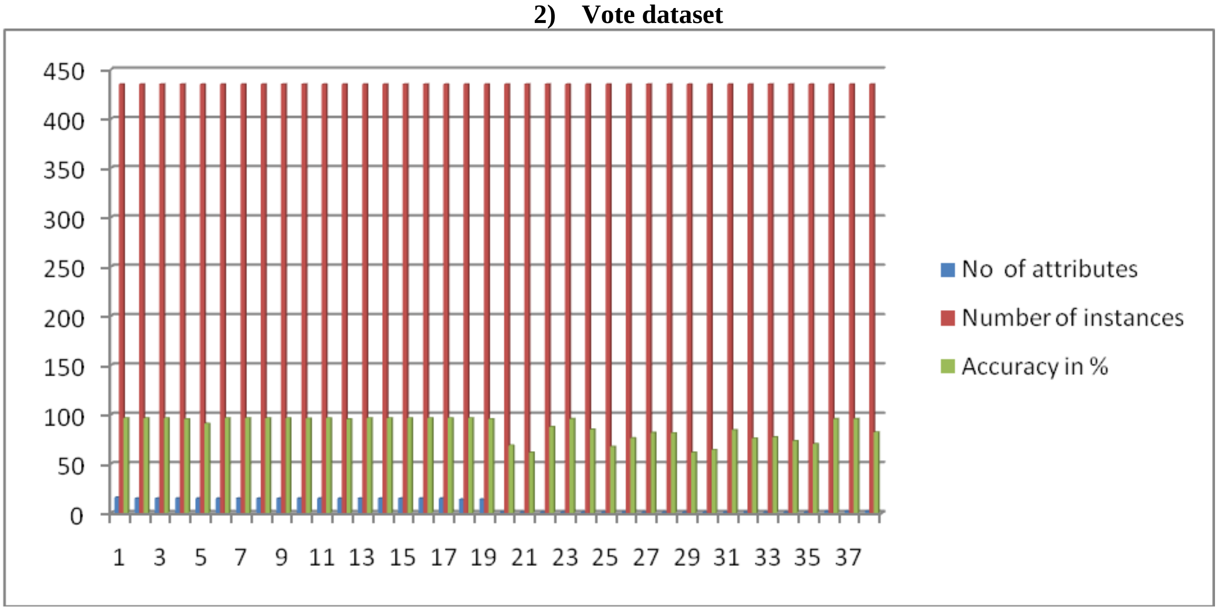
1) Weather dataset



Best attribute: Temperature

Here we have selected temperature is the best attribute which alone gives the accuracy of 64.28 % which is far better than the accuracy provided by the four attribute taken together 50%

2) Vote dataset



Best attribute: physician-free-freeze

For the vote dataset, which consist of 16 attributes which together gives accuracy of 96.32%, after applying the algorithm we have found out that even single attribute physician-free-freeze gives the accuracy of 95.63%, which is very efficient. Hence for this dataset it can be considered as the best attribute.

IV. Conclusion

Thus in this paper we have tried to select best significant feature of data set using repeatedly applying the decision tree classification algorithm J48 over the dataset and found that for weather dataset temperature is the best attribute which alone gives the accuracy of 64.28 % which is far better than the accuracy provided by the four attribute taken together 50% .For the vote dataset, which consist of 16 attributes which together gives accuracy of 96.32%, after applying the algorithm we have found out that even single attribute physician-free-freeze gives the accuracy of 95.63%, which is very efficient. Hence for this dataset it can be considered as the best attribute. Future work will focus on analysis on the feature selected with various other techniques like principle component analysis, genetic algorithms.

References

[1]. Olivier Henchiri , Nathalie Japkowicz. “A Feature Selection and Evaluation Scheme for Computer Virus Detection, Proceedings of the Sixth International Conference on Data Mining “, (ICDM'06).

[2] M. Dash, H. Liu, Feature Selection for Classification, Intelligent Data Analysis 1 (1997) 131-156.

[3] Frank,A. & Asuncion ,A. (2010).UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].Irvine,CA: University of California, School of Information and Computer Science.

[4] Data mining book – kamber

[5] V.Vaithiyanathan, K.Rajeswari, Swati Tonge,, “ Improved Apriori algorithm based on Selection Criterion”, IEEE Conference ICCIC Dec 18-20, 2012, Coimbatore. Catalog Number: CFP1220J-ART ISBN: 978-1-4673-1344-5

[6] Machine Learning Group at the University of Waikato:http://www.cs.waikato.ac.nz/ml /weka/