www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-6, Issue-6, December 2018
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Efficiency of Data Mining Algorithms Used In Agnostic Data Analytics Insight Tools

A.Jenita Jebamalar

Dept. Computer Science, St.Thomas Matric Higher Secondary School, Thoothukudi, India

Received: 03/Dec/2018, Accepted: 21/Dec/2018, Published: 31/Dec/2018

Abstract - Insights are the results of the analytics with various parameters like customer demographics, gender, age, behavior, interests, etc. The objective is to predict which product the customers are least likely and most likely to buy. The result of the analytics is the insights which are provided in the form of tables, charts and graphs. In the technology world, the term agnostic means that the tools are not restricted to a specific systems and it works with various systems rather than being designed for a single system. Agnostic data means that it does not comes from a specific source. In machine learning, feature selection is used to reduce the properties of the class variables by removing the redundancy from the dataset. The goal of this research work is compare and find the efficiency of various data mining algorithms used in analytics insight tools. Dataset is collected from an analytics of a website for the listed algorithms. Data mining utilizes algorithms, statistical analysis and even artificial intelligence to extract data from huge data sets into an apprehensible form. The future work will be the implementation of the selected algorithm in the data analytics insight tool.

Keywords: *Agnostic, Insights, Feature Selection Algorithms, Data Analysis, Data Discovery.*

I. INTRODUCTION

In Data Analytics Insight tool it is really an interesting area of research in finding out the efficient data mining techniques in data discovery and extracts useful information. A website's Google analytics dataset has been used. After the data collection, data cleaning, user identification, pattern discovery and analysis has been done. Different data mining classifiers are identified and evaluated for accuracy and precision to identify which data mining technique is suitable for the analytics insight tool. Most used classifier in this research is Naïve Base Classifier. This paper work is mainly focused on improving the results in data analytics and insights using the right data mining techniques. Finally the classification tasks are conducted using efficient classifiers like Radial Basis Function, K-Nearest Neighbor and Artificial Neural Network to estimate the data analytics and insight performance. This paper analyzes the various classification algorithms and their performances are equated using WEKA tool and results are discussed. The algorithm resulted of this research work can be used in the data analytics insight tools.

This paper is integrated as follows. Section 2 talks about background of the study. Section 3 discusses various feature selection techniques used. The statement of the problem is furnished in Section 4. The details of the dataset generated for the study is provided in the Section 5. The observational evaluation and comparative analysis are explained in Section 6 and Conclusion for the work is mentioned in Section 7. Lastly, vital references are mentioned in Section 8.

II. BACKGROUND

Feature selection is a vital part in machine learning. It refers to the process of finding the useful inputs and extracting the useful information or features from the input data. It is very difficult to discover the meaningful patterns if the data is redundant. Most of the data mining algorithms need a larger training data set if the dataset is high-dimensional. Machine learning algorithm generally scores columns and validates the dataset. Feature selection is always been performed before the model is trained. Each algorithm has its own set of default techniques for applying feature reduction. Still we can manually set parameters to influence the feature selection behavior. In the automatic feature selection, a score is calculated for each attribute, and only the attributes that have the best scores are selected for the model. Multiple methods are provided to calculate for calculating the scores and the method applied are depends on the factor like, the algorithm used, data type of the attributes and parameters set on the model. Feature selection is applied to data inputs and predictable attributes. Feature selection has no effect on storage of the mining structure where in it affects the columns used in the model. For discrete data, Shannon's entropy and Bayesian scores are used. For continuous columns, the interestingness score is used to assess the input columns and to ensure the consistency. The measure of interestingness is entropy based. The entropy of a particular

attribute is compared to the entropy of all other attributes as Interestingness (Attribute) = - (m – Entropy (Attribute)) * (m – Entropy (Attribute)). Central entropy m is the entropy of the entire feature set. By subtracting the entropy of the target attribute from m, we can assess information size that the attribute provides. Shannon's entropy assesses the uncertainty of a random variable for a particular outcome. The formula to calculate Shannon's entropy is

$$H(X) = -\sum P(xi) \log(P(xi))$$

The above method is used for discrete attributes. Bayesian network is a acyclic graph of states and transitions between states, meaning that some states are always prior to the current state, some states are posterior, and the graph does not repeat. Bayesian networks allow the use of prior knowledge. However, for the algorithm design, the question of which prior states to use in calculating probabilities of later states is important for performance and accuracy. The K2 algorithm for learning from a Bayesian network is often used in data mining which was developed by Cooper and Herskovits. It is actually scalable and can analyze multiple variables which require ordering on variables used as input.

III. STATEMENT OF THE PROBLEM

There are various tools available to pull the data of a website and provide the analytics and insights as table and charts. In this research, the efficiency of various feature selection algorithms are evaluated using the Google Analytics dataset of a magician website generated for this study. The proposed study has compared several feature selection techniques to find its efficiency. The efficiency resulted is by using the measures of error and accuracy parameters. The dataset used in this study included the demographic details of the visitors, their gender, location and other behaviors. The goal of this study is to find out the efficient algorithm for the analytics and insight tool. Therefore it could be useful the companies that create the data and insight tools.

IV. RESEARCH METHODOLOGY

In this section various feature selection algorithms are discussed. Further the feature selection parameters and modeling flags are illustrated.

A. Naïve Bayes Algorithm

In this algorithm, the methods of analysis used are Shannon's entropy, Bayesian with K2 prior and Bayesian Dirchilet with uniform prior. This algorithm accepts only discrete attributes. So it cannot use the interestingness score.

B. Decision Trees Algorithm

This algorithm uses the analysis methods like Interestingness score, Shannon's entropy, Bayesian with K2 prior and Bayesian Dirichlet with uniform prior. If any of the columns contain non-binary continuous values, the interestingness score is used for all columns, to ensure the consistency. Otherwise, default feature selection method is used, or the method that specified while the model is created.

C. Neural Network Algorithm

Interestingness score, Shannon's entropy, Bayesian with K2 prior and Bayesian Dirichlet with uniform prior are used as the method of analysis. As long as the data contains continuous columns, this algorithm use both Bayesian and entropy-based methods.

D. Logistic Regression Algorithm

The methods of analysis used are Interestingness score, Shannon's entropy, Bayesian with K2 prior and Bayesian Dirchilet with uniform prior. It is actually based on the neural network algorithm but we cannot customize logistic regression models to control feature selection behavior. Hence feature selection always default to the method that is most appropriate for the attribute.

E. Clustering Algorithm

The analysis method used is Interestingness score. It uses the discrete data. The score of each attribute is calculated as a distance and is represented as a continuous number and the interestingness score must be used.

F. Linear regression Algorithm

This algorithm can only use the interestingness score, because it only supports continuous columns.

FEATURE SELECTION PARAMETERS

Each algorithm always has a default value for the number of inputs that are allowed, but we can override this default and specify the number of attributes. This section lists the parameters that are furnished for managing feature selection.

MAXIMUM_INPUT_ATTRIBUTES

If a model contains more columns than the number that is specified in the MAXIMUM_INPUT_ATTRIBUTES parameter, the algorithm dismisses any columns that it calculates to be uninteresting.

MAXIMUM_OUTPUT_ATTRIBUTES

Likewise, if a model contains more predictable columns than the number that is specified in the MAXIMUM_OUTPUT_ATTRIBUTES parameter, the

algorithm ignores any columns that it calculates to be uninteresting.

MAXIMUM_STATES

If the model contains more cases than are specified in the MAXIMUM_STATES parameter, the least popular states are grouped together and treated as missing. When any one of these parameters is set to 0, feature selection is turned off, affecting processing time and performance.

Apart from these methods for feature selection, we can improve the ability of the algorithm to identify or promote meaningful attributes by setting modeling flags on the model or by setting distribution flags on the structure.

MODELING FLAGS

Modeling flags are used to provide additional information to a data mining algorithm about the data that is defined in a case table. The algorithm uses this information n to build a more accurate data mining model. Few modeling flags are defined at the level of the mining structure, whereas others are defined at the level of the mining model column. For example, the NOT NULL modeling flag is used with mining structure columns. We can define additional modeling flags on the mining model columns, depending on the algorithm we use to create the model.

LIST OF MODELING FLAGS

A. NOT NULL

This indicates that the values for the attribute column should never contain a null value. An error will result if it encounters a null value for this attribute column during the model training process.

B. MODEL_EXISTENCE_ONLY

This indicates that the column will be treated as having two states: Missing and Existing. If the value is NULL, it is treated as Missing. The MODEL_EXISTENCE_ONLY flag is applied to the predictable attribute and is supported by most algorithms. In implementation, setting the MODEL_EXISTENCE_ONLY flag to True changes the representation of the values such that there are only two states: Missing and Existing. All the non-missing states are combined into a single Existing value.

A distinctive use for this modeling flag would be in attributes for which the NULL state has an implicit meaning, and the explicit value of the NOT NULL state might not be as important as the fact that the column has any value. For example, a [Date Contract Signed] column might be NULL if a contract was never signed and NOT NULL if the contract was signed. Therefore, if the purpose of the model is to predict whether a contract will be signed, you can use the MODEL_EXISTENCE_ONLY flag to ignore the exact date value in the NOT NULL cases and distinguish only between cases where a contract is missing or existing.

C.REGRESSOR

This indicates that the column is a candidate for used as a regressor during processing. This modeling flag is defined on a mining model column, and can only be applied to columns that have a continuous numeric data type. If we set the REGRESSOR modeling flag on a column, we are indicating to the algorithm that the column contains potential regressors. The actual regressors that are used in the model are determined by the algorithm. A potential regressor can be dimissed if it does not model the predictable attribute.

D. Regressors in Linear Regression Models

Linear regression models are based on the Decision Trees algorithm. Any decision tree model can contain a tree or nodes that represents a regression on a continuous attribute. Hence, in these models we do not need to specify that a continuous column represents a regressor. Decision Trees algorithm will partition the dataset into regions with meaningful patterns even if we do not set the REGRESSOR flag on the column. The difference is that when we set the modeling flag, the algorithm will try to find regression equations of the following form to fit the patterns in the nodes of the tree.

$$a*C1 + b*C2 + ...$$

Then, the sum of the residuals is calculated, and if the deviation is too much great, a split is forced in the tree. For example, if we are predicting customer purchasing behavior using Income as an attribute, and set the REGRESSOR modeling flag on the column, the algorithm would try to fit the Income values by using a standard regression formula. If the deviation is too great, the regression formula is deserted and the tree would be split on some other attribute. The decision tree algorithm would then try fit a regressor for income in each of the branches after the split.

FORCE_REGRESSOR parameter is to guarantee that the algorithm will use a particular regressor. This parameter can be used with the Decision Trees algorithm and Linear Regression algorithm.

V. EXPERIMENTAL DATA

A website's analytical dataset was generated based on the demographics and behavior. Two months of the website visitor's analytical data was generated and used for this study. Irrespective of the algorithm that was used, mining model content is presented in a standard structure. The content of each model is presented as a series of nodes. A

node is an object within a mining model that contains information about a portion of the model. Nodes are arranged in a hierarchy. The exact arrangement of nodes in the hierarchy depends on the algorithm used. For example, when we create a decision trees model, the model can contain multiple trees, all connected to the model root. When we create a neural network model, the model may contain one or more networks, plus a statistics node. The first node in each model is called the root node, or the parent node. Every model has a root node (NODE_TYPE = 1). The root node typically contains some data about the model, and the number of child nodes, but some additional information about the patterns discovered by the model. Depending on the algorithm used to create the model, the root node has a varying number of child nodes. Child nodes have different meanings and contain different content, depending on the algorithm and the depth and complexity of the data. Following are the additional content included in the node.

- Statistics, such as standard deviation, mean, or variance.
- Rule definitions and lateral pointers.
- Count of cases in the training data which supports a particular predicted value.
- Formulas and coefficients.

Table 1: Reduction of Attributes using Feature selection algorithm. WEKA software tool was used.

City Attribute Data Type: Text	Existing Users Attribute Data Type: Numeric	New Users Attribute Data Type: Numeric	Session Attribute: Data Type: Numeric (Minutes)	Duration Attribute Data Type: Time: HH:MM:SS
Chennai	20	20	20	1.65
Sao Paulo	5	5	5	1.00
Gold Coast	1	1	1	1.00
Fortaleza	1	1	1	1.00
Dourados	1	1	1	1.00
Joao Pessoa	1	1	1	1.00
Recife	1	1	1	1.00
Teresopolis	1	1	1	1.00
Florianopolis	1	1	1	1.00
Barueri	1	1	1	1.00
	49	48	49	1.29

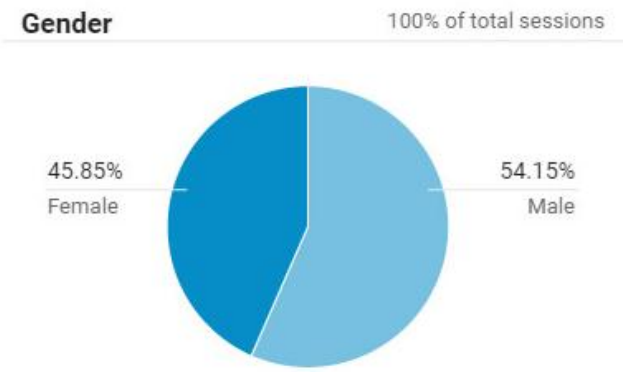

Figure 1: Gender Results

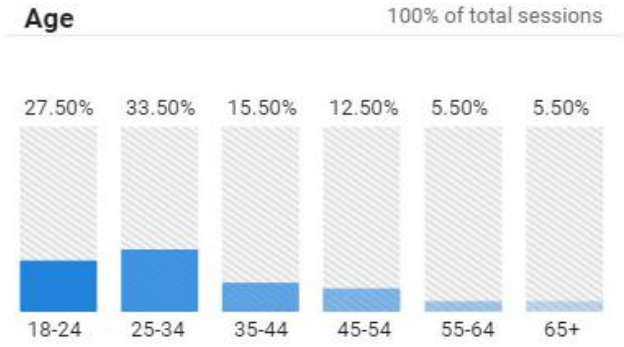

Figure 2: Age Results

Figure 3: Interest Category Results

VI. RESULTS AND DISCUSSIONS

The efficiency of this model is mainly depends on selection of best Attributes from the list of attribute used in website's analytical data set. The present method concentrates on different algorithm used in data preprocessing. The effectiveness of the algorithm is demonstrated in terms of different measures. For assessing the excellence in this,

Receiver Operating Characteristics (ROC) value can be used.ROC value is the representation of the tradeoff between the false positive and false negative rates. F-Measure, that is another measure for evaluating the effectiveness, is the sympathetic mean of the precision and recall. The evaluation measures with variations of ROC values and F-Measure are the output from an Open Source Data mining tool WEKA.

PERFORMANCE EVALUATION:

F-measure and area under Receiver Operating curve (ROC) are the metrics used for the proposed model evaluation. True positives are the correctly classified positive case and false positives are incorrectly classified positive cases. True negatives are correctly classified negative cases and false negatives are incorrectly classified negative cases.

The below graph indicates that the classifier Naïve Bayes could attain a highest ROC value and Bayes Net had second highest value. So, we attain that Naïve Bayes has the optimal dimensionality in the website analytics dataset.

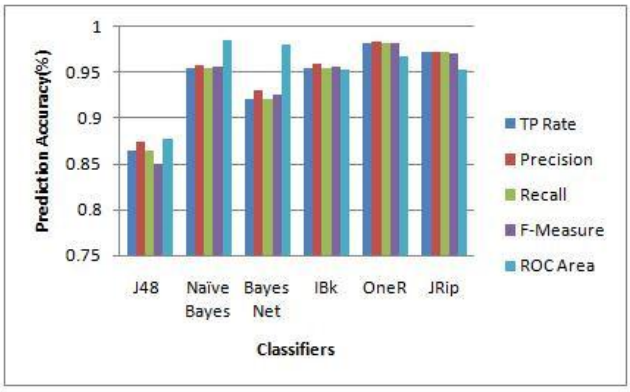

Figure 4: Classification Results

VII. CONCLUSION

The aim of this research work is to identify the better data mining algorithm while generating the insights for the agnostic analytics data. This research work should very much useful for the agnostic data analytics tool creators. The Naïve Bayes' classifier gave a very high accuracy and identified as the better data mining algorithm in this study. The future work will be handling the large datasets of multiple web analytics.

REFERENCES

[1]. Aldekhail M, *"Application and Significance of Web Usage Mining in the 21st Century: A Literature Review",* International Journal of Computer Theory and Engineering, Vol.8, Issue.1, pp. 41-47,2016..

[2]. Tarik A. Rashid, "*Improving on Classification Models of Multiple Classes through Effectual Processes"*, International Journal of Advanced Computer Science and Applications, Vol.6, Issue.7, pp.55-62, 2015.

[3]. Forman, George, *"An extensive empirical study of feature selection metrics for text classification"*, The Journalof machine learning research, Vol.3, pp.1289-1305, 2003.

[4]. Egozi, Ofer, EvgeniyGabrilovich, and Shaul Markovitch, *"Concept-Based Feature Generation and Selection for Information Retrieval"*, AAAI, Vol.8, pp. 1132-1137, 2008.

[5]. Russel, Stuart, and Peter Norvig, "Artificial Intelligence: A Modern Approach ", EUA: Prentice Hall, 2003.

[6]. Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok, *"Adaptive intrusion detection: A data mining approach"*, Artificial Intelligence Review, Vol.14, Issue.6, pp. 533-567, 2000.

[7]. Cooley R., Mobasher B. and Srivastava J, *"Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems"*, Vol.1, Issue 1, pp. 5-32, 1999.

[8]. Vaibhav P. Vasani, Rajendra D. Gawali, *"Classification and performance evaluation using data mining algorithm"*, International. Journal of Innovative Research in Science, Engineering and Technology, Vol.3, Issue 3, pp.10453-10458, 2014.

Authors Profile

Mrs.A.Jenitta Jebamalar pursued her Bachelor of Computer Science in St.Mary's College, Tuticorin. She also done her Master of Science in both Information Technology in St.Xavier's College, Palayamkottai and Computer science in Annamalai University. She completed her Master of Philosophy in Computerscience in Manonmaniam Sundaranar University, Tirunelveli. She started her teaching experience from 2006 in Bishop Caldwell College, Tuticorin and moved on to Women's Christian College, Chennai to cover her 10 years of her service since 2018.She is persuing her Phd in Computer science in the field of DATA MINING and currently working as Computer Instructor, St.Thomas. Mat.Hr.Sec.School, Tuticorin.She also published three more papers in National Seminars and Conferences.

