Journal of Engineering and Science Research, 2 (1): 11-17, 2018
e-ISSN:2289-7127

DOI:10.26666/rmp.jesr.2018.1.3

Performance Evaluation of Different Classifier for Big data in Data mining Industries

Eloanyi Samson Amaechi and Prof. Hai Van Pham

Hanoi University of science and Technology (HUST), No 1.Dai Co Viet.Hanoi Vietnam

ABSTRACT: Data mining is the set of computational techniques and methodologies aimed to extract knowledge from a large amount of data, by using sophisticated data analysis tools to highlight information structure underlying large data sets. Data scientist and data engineer are facing big challenges today in society because of global increases in the dataset in the industries and sector today. Machine learning methods represent one of these tools, allowing, not only data management but also analysis and prediction operations. Supervised learning, a kind of machine learning methodology, uses input data and products outputs of two types: qualitative and quantitative, respectively describing data classes and predicting data trends. Classification task provides qualitative responses whereas prediction or regression task offers quantitative outputs. In this paper, an attempt has been made to demonstrate how big data can be analyzed, classified and predicted using weka tool in industries.

Keywords: Dataset, Weka Tool, Supervisor Learning, Unsupervisior Learning, Classification, Classifies and Predication.

INTRODUCTION

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from databases/data warehouses. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form, which is easily comprehensible to humans [1]. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help user focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining goes beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They source databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line .After the success of the data processing results every step is determined from the choice of data and ending with an explanation of the anomalies in the results. This paper is focus on the characteristics classifiers and how to analyzed, classified, predict, optimized and improved the nature for the developer work in the future. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can

Corresponding Author: Prof Hai Van Pham, Hanoi University of sciences and Technology (HUST), No 1. Dai Co Viet..Hanoi Vietnam, samsonj688@gmail.com

then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps [2].KDD means knowledge discovery in database**.**

I WEKA

Weka workbench is a collection of machine learning algorithms and data preprocessing tools. It is designed to quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools.Weka is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code [3].Originally, weka tool was designed for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

Weka techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java. Database connectivity can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using weka. Another important area that is currently not covered by the algorithms included in the weka distribution is sequence modeling [4].Weka tool support all the standard data especially data preprocessing, clustering, classification, regression, visualization, and feature selection.

II. DATA PROCESSING

The primary available dataset such as FPT Software is developers dataset, with basic information contained about working hours, the nature of work the developer want to undergo in future, which required the knowledge of design and developing the database for the information. The aim is to determine tasks status of developer: Notsure, NotWork, Not Late and Late, which help developer to classify and predicted the outcome of tasks in the future. Classification and prediction enable the developer to determine which tasks finished within the hours, which tasks can't finish within the hours, which tasks can be workable since it is big data. The dataset contained the name of tasks, date for issued; hours to started, hours to finished and the name for the developer the task is issued. The data processing in this experiment start with the following points:

1. Extraction of dataset from excel format to CVS file format and CVS format to JSON format. A comma separated values (CSV) file contains different values separated by a delimiter, which acts as a database table or an intermediate form of a database table. In order to understand the complete nature of the dataset used for experiment, we convert from the excel format to CSV format, because CSV format is used to store tabular data, such as spreadsheets or database. JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. JSON is a text format that is completely language independent but uses conversions that are familiar to programmers of the C-family of languages, including C, C++ C# Java, JavaScript, Perl, Python and many others [5]. Dataset are big data and there are some reasons of converting from CVS format to JSON format. there are:

- Easy to view and readable of dataset
- .Easy to view the data log easily and translate to any format when working in the data**.**

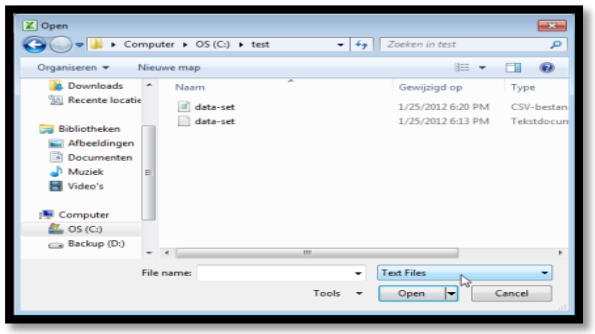

Fig 1 *Dataset extraction from Excel, CSV and JSON format*

3. Creating of database for dataset: The database is designed in SQL -2008 database management system to store the collected data. Structured Query language is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS) or for stream processing in a relational data stream management system (RDSMS). In this project we created a database for the data log in order to have content of another layer of legal protection, for flexibility and easier management of website, makes navigation and searching easier for both user and administrator inevitable. The data is formed according to the required format and structures. Database was designed to maintain and manage the dataset which is:

- Easily accessed, managed and updated.
- Easy to capture and analyzed the dataset.

4. Tagging dataset (Categorizing and Tagging words): the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context in order to prepare a dataset for process and to remove some unwanted data inside the dataset. There are some reasons why tagging of dataset is necessary to carry out before the data will be accepted in Weka environment, these are

- Easy to analyzed the data log which means there are many misinformation in dataset
- To understand the nature for dataset when applied in software tool

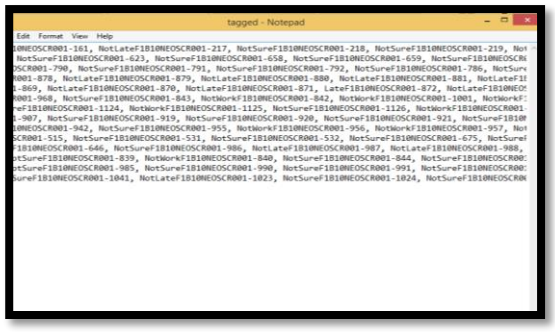

Fig 2 *Dataset Tagging processing*

4. Database is designed in SQL -2008 database management system to store the collected data. The data is formed according to the required format and structures. Further, the data is converted to ARFF (Attribute Relation File Format) format t3 process in WEKA. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. After processing the ARFF file in WEKA the list of all attributes, statistics and other parameters can be utilized as shown in Figure 3a and 3b

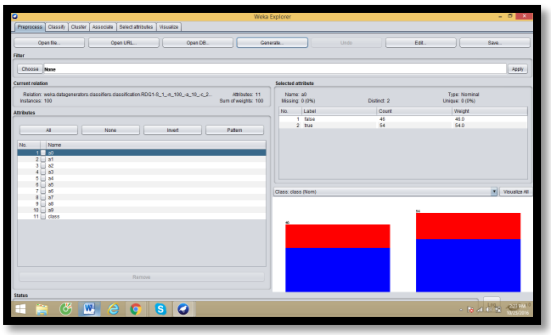

Fig 3 a. Processed ARFF file in WEKA environment

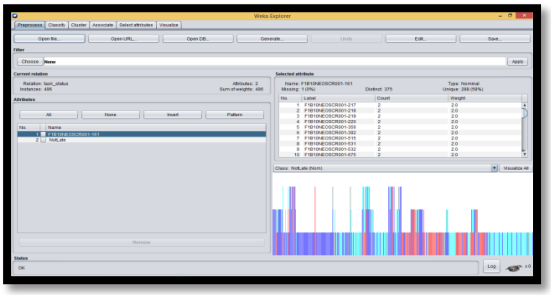

Fig 3b Processed ARFF file in WEKA environment

III: METHODOLOGY

FPT software data contained information about work tasks status for developer Notsure, NotWork, Not Late and Late, which the developer must classified and predict the out of the work future. Meanwhile, the data explained the information about the optimization of time during work hour future. In this paper we focus on how to analyzed, classified, predict, optimized and improved the nature for the developer work in the future. Data classification is process of organizing data into categories for its most effective and efficient use. In other to determine and categories datasets into difference machine leaning models. A well-planned data classification system makes essential dataset easy to find and retrieve. We applied Weka software tools to build classification models in machine leaning so as to run on machine learning algorithms. WEKA (The Waikato Environment for

Knowledge Analysis) software is used. WEKA workbench aids to apply machine learning techniques for myriads of real world problems. The WEKA machine learning workbench provides an environment for automatic classification, regression, clustering and common data mining problems in bioinformatics research. It has a user friendly graphical interface to compare the various algorithm results [6]. In this paper, we use classification and data analysis of FPT Software dataset. Naïve Bayes classifier, SMO (support vector machine), J48, Decision Tree, ZeroR (Instance based classifier), have been used in order to analyze the results.

MACHINE LEARNING ALGORITHMS

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets methods at the intersection of machine learning such as in FPT software data set. The main aim of the data mining process is to retrieve the data from data set, and transform into more meaningful form with the help of the algorithms. In this paper, machine learning algorithms developed for data mining is used. These five algorithms are used to analyze the results. The algorithms are Naïve Bayes classifier, SMO (support vector machine), J48, Decision Tree, ZeroR (Instance based classifier). The weka workbench assists to retrieve the Software for running the algorithms. Thus, we can easily realize the difference between the algorithm results. To sum up, the best classification on the FPT Software data set is understandable. The following are the classifiers use in analyzing FPT software data and it performance. There are:

Naïve Bayes Classifier: Naïve Bayes is a statistical learning algorithm that applies a simplified version of Bayes rule in order to compute the posterior probability of a category given the input attribute values of an example situation. Prior probabilities for categories and attribute values conditioned on categories are estimated from frequency counts computed from the training data. Naïve Bayes is a simple and fast learning algorithm that often outperforms more sophisticated methods. It is a supervised type of learning algorithm and literature said that it is extremely fast as compare to sophisticated methods [7].The Bayesian classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve, diagnose and predictive problems. The performance of this classifier is 82% for current corpus

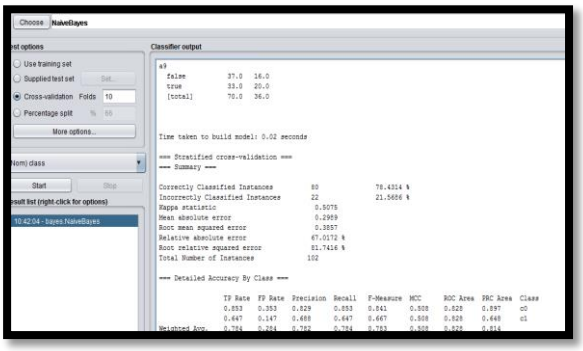

Fig. 4 *Dataset Extraction of Naïve Bayes Classifier.*

SMO (Support Vector Machine): SMO implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels. Sequential minimal optimization (SMO) is an algorithm used to solve the quadratic programming problem which arises during training of support vector machine. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier.).The performance of this classifier is 74% for current corpus

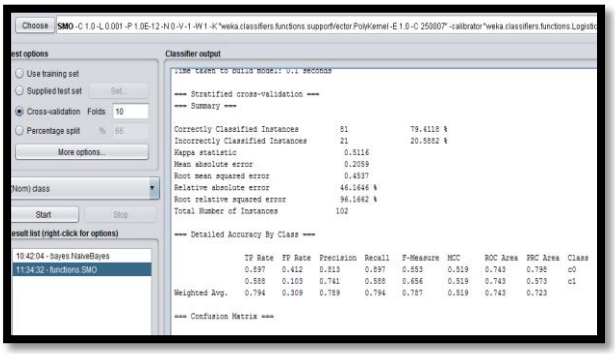

Fig 5. *Dataset Extraction of SMO classifier.*

J48: J48 is a classifier using divide and conquer strategy for decision making. Divide and conquer strategy is solved by recursive procedure or recursive tree procedure based on this; it will create desiccation tree for classification of eye state whether it is open or closed. A decision tree helps decision support system and it uses tree like structure. It is used to learn a

classification function it requires dependant variable and independent variable for classification. Decision tree have many advantages over other classification algorithms because it is having ability to work with variety of input data types (nominal scale, numeric scale and textual data) this classifier is tested using weka. The performance of this classifier is 81% for current corpus

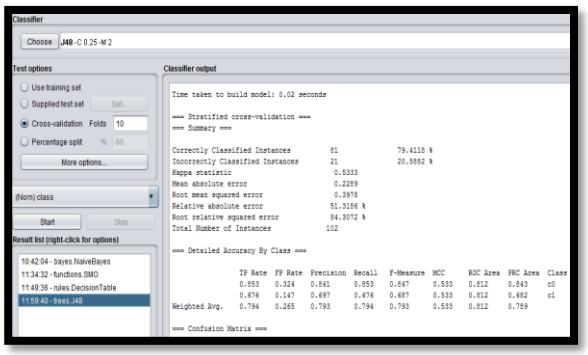

Fig 6 *Dataset Extraction of J48Classifier.*

Decision Tree: Leo Breiman and Adele Cutler introduced random tree classification algorithm. The beauty of this algorithm is that it works with both regression and classification problems. Decision tree doesn't require accuracy estimator [8]. Decision tree algorithm built multiple decision trees randomly. Decision tree analysis on J48 algorithm is applied to Weka. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. The performance of this classifier is 85% for current corpus

ZeroR (Instance based classifier): Is an instance-based classifier that is the class of a test instance and based upon the classes of those training instances similar to it, as determined by some similarity function. The dataset includes 100 instances of two class or labels, true or false. Data division is the process of division of dataset in the weka environment. In classification, the dataset is divided into training set and tested set with the percentage of 60% for training set and 40% for testing set. The performance of this classifier is 45% for current corpus.

TABLE I: VARIOUS MEASURE OF CLASSIFIER

Classifier	TP rate	FP rate	Precision	Recall	F-measure	ROC Area	Accuracy	Error rate
Naive Bayes	0.78	0.28	0.78	0.78	0.78	0.82	82%	18%
SMO	0.79	0.30	0.78	0.79	0.78	0.74	74%	26%
J48	0.79	0.26	0.79	0.79	0.79	0.81	81%	19%
Decision Tree:	0.79	0.25	0.79	0.79	0.79	0.85	85%	15%
ZeroR	0.54	0.54	0.30	0.54	0.38	0.45	45%	55%

IV EXPERIMENT

In proposed work the developer FPT dataset for experiment is taken and the process diagram is shown in Table. 1.

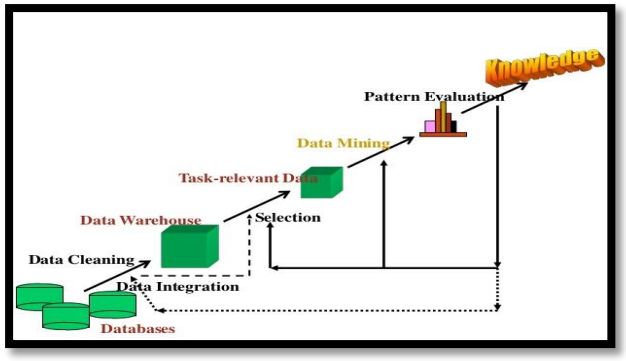

Fig 7 *Dataset process diagram*

After data extraction outlier detection is performed for purifying the extracted data from different electrode positions. After this feature estimation is carried out and some features are selected from all features. Then it performs feature classification based on different types of classifier such as Tree based, rule based, Bayesian based, Function based etc. The various statically measures for classifier is evaluated and different measure like TP Rate, FP Rate, Precision, Recall, F-measure,. Accuracy of Instance Based classifier is compared with other classifier and it is TPR (True Positive Rate): It is also known as sensitivity which measures the actual positive classified instances in binary classification.

TPR = TP / (TP + FN)

where:
TP is True positive,
FN is False Negative.
FPR (False Positive Rate): It is also known as false alarm ratio. It measures the expectancy of false positive ratio of instances classified.

FPR = FP / (FP + TN)

where:
FP is false positive
TN is True Negative.
Precision: The proportion of predicted positive which are actual positive. It is the fraction of retrieved instances that are relevant in recognition with binary classification. It is also known as positive predicted value.

Precision = TP / (TP + FP)

where
TP is True positive
FN is False Negative
F-measure: It is harmonic mean between Precision and Recall and also known as F-score. It considers both the precision and the recall of the test to compute the score: precision is the number of correct positive results divided by the number of all positive results, and recall is the number of correct positive results divided by the number of positive results that should have been returned. The F-score can be interpreted as a weighted average of the precision and recall.[9]

F-measure = 2(Precision × Recall) / (Precision + Recall)[10] .

ROC curve (Receiver Operating Characteristic curve): In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. It is a 2D curve parameterized by one parameter of the classification algorithm. In a ROC curve the TP rate (sensitivity) is plotted in function of FP rate for different cut-off points of parameter [11] The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR).

CONCLUSION

Knowledge of big data extraction from database has become vital importance since the data from industries increase from megabyte to terabytes very day, it become one of the key process which ever organization face in their sectors. In this paper, Decision Tree classifier gave the best accuracy when compared with different classifier 85.0% followed by Naive Bayes with 82.0% while base meta is 45.0%. So the instance based classifier is better choice to predict developer tasks in future whether to work or not work. This analysis help organization and industries to predicate their future work base on the data available.

ACKNOWLEDGMENTS

Authors are grateful to the department of MSE, for providing all the basic data and Weka tool for providing such a strong tool to extract and analyze knowledge from database.

REFERENCES

[1]kochetov Vadim, Overview of different approaches to solving problems of data mining, vol.123, 2018 pages 234-239.
[2]Mridu Sahu, N. K. Nagwani, Shrish Verma, and Saransh Shirke,Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal,Vol. 1, No. 2, September 2015
[3] Jiawei Han and MichelineKamber, Data Mining Concepts and Techniques, 2nd ed., Morgan Kaufmann publishers, SanFrancisco, 2006.
[4] George M. Marakas, Modern Data Warehousing, Mining, and Visualization, Pearson Education, New Delhi, 2005.
[5] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Pearson Education, New Delhi, 2009
[6]Bishnu Prasad De, R. Kar, D. Mandal, S. P. Ghoshal, "Optimal selection of components value for analog active filter design using simplex particle swarm optimization", International Journal of Machine Learning and Cybernetics August 2015, Volume 6,
[7] Sunita B Aher, Lobo LMRJ, Data Mining in Educational System using Weka, International Conference on Emerging Technology Trends (ICETT), Proceedings published by International Journal of

Computer Applications® (IJCA) Number 3, 2011, pp-20-25.

[8] Michael J.A. Berry and Gordon S. Linoff, Data Mining Techniques, 2nd ed., Wiley Publishing Inc., USA, 2004

[9] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994. Machine Learning, Neural Statistical Classification, Ellis Horwood.

[10] Witten, I.H. & Frank, E. 2000. Weka Machine Learning Algorithms in Java, in Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, pp. 260- 320

[11] C. Carmelo, P. Montalto, M. Aliotta, A. Cannata, and A. Pulvirenti. ―Similarity measures and dimensionality reduction techniques for time Series data mining,‖ Advances in Data Mining Knowledge Discovery and Applications, 2012