International Journal of Management, IT & Engineering
Vol. 8 Issue 9, September 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

COMPARING DIFFERENT CLASSIFICATION TECHNIQUES USING DATA MINING TOOLS

Bhawana Mathur[*]

Manju Kaushik*

Abstract

The model has been developed by classification methods such as Chidambar and Kemerar Matrix Suit, and by LoC Metric, ADT, Decision Table, Hyper Pipe, and Naive Bayes. Results depend on 05 open sources Java software with diverse fields such as databases, formats, and protocols, games and entertainment, multimedia, science and engineering. These have been used to test the total sample of 8 characteristics (one for seven for input and one for output), with 3518 examples. The prototype is then evaluated using two specific performance parameters, precision, and F-measurement. In this investigation, we used classification strategies that use the data mining tool.

Keywords: ADTree; Decision Table; Hyper Pipes; CK Metrics; Object-Oriented Software; Performance Measurement.

[*] **Department of Computer Science & Engineering, JECRC University, Jaipur (India)**

1. Introduction

In this paper, "Comparison of different classification techniques using data mining tool Weka", the authors used MATLAB with the WEKA tool. The purpose of this paper is to measure and test the specific classification techniques Naïve Bayes, special ADT, decision table, hyper pipes (Fan *et al.*, 2008), (Wahbeh *et al.*, 2011), (Dong *et al.*, 2005). Analyzing the validity of this evaluated software metric, it has inspired us to estimate that many custom, as well as object-oriented metrics, then provide designers, coders, and valuable data to analysts as well. This proposed work does not display around a huge help in the performance of software metrics for software quality assessment, although there is also a specialty to efficiently survey the structure through these metrics. Veka has developed scholars with modern analysts with very mainstream. Generally, education is also used for dedication. For WEKA, Regression, Classification, Clustering, Association Rules, Visualization, as well as Data Pre-Processing (http://www.cs.waikato.ac.nz/ml/weka/), (Goebel and Grunwald, 1999) Equipment included. Data mining is a "decision-support" technique in which data design is searched for data. This strategy can be used on some types of data. Verification models require verification of cross-verification. The model's performance has been evaluated by 10-fold cross-verification (Braga-Neto and Dougharty, 2004) (Sharma and Sahni, 2011). For this research, the data sets (Sharma and Sahni, 2011) are set parallel to the execution of various classification processes. Open source data has been used to test 3518 examples as well as to test 8 specifications (7 for input and 1 for output) (Kaur and Lion, 2016). Comparison of software packages from mining in this paper contains information about study classification techniques. This article is useful in relation to the reform of administrative decisions with open-source machine learning software (Hornik *et al.,* 2009) (Altintus *et al.,* 2004), (Hall *et al.,* 2009). The purpose of this research is to find the best machine learning (ML) algorithm for problem (Khoussainov *et al*., 2004). In this paper, Weka was evaluated on the performance data of the machine learning model. Partition and K-fold cross-verification (Rodriguez et al., 2010) Four unique test options and each have been tested to use, focuses on the implementation of the issues of metrics, classification and regression (Stumpf *et al*., 2009).

2. Related Works

In this section, we highlight the contribution of some researchers who have made valuable contribution in our field. At the present time, many people are working with machine learning processes (Caruana, and Niculescu -Mizil, 2006). On accuracy, different data sets and specific parameters are highlighted for Multilevel Perceptron, J. 48, to perform three processes like Naive Bayes (Hall *et al*, 2009). Witten and Frank are a comprehensive source of data available in Data Mining (2005) and have been included in User Manual Software Distribution (Hall et al, 2009), (Witten et al, 2016). Generally, supervised statistical learning strategies such as many machine learning strategies (Dietterich et al., 1997) have been used. In the supervised education, the present requirements (Dietrich, 1998), using the information, meet the ideal requirements. Some system changes are accessible for classification techniques such as the regression model. All machine learning techniques are used (Dietterich *et al.,* 1997). Prior to the initial use of preset classes (Dietterich, 1998), prototype requirements and information have been used to learn learning techniques. It affects a prototype, which is used to label / classify analysis examples. Class label standards are considered unknown. Four distinct classifications, especially the Naive Bayes (NB), Alternative Decision Tree (ADT), Hyper pipe are compared to the Decision Table. These classification algorithms are known as high-performance change forecasts (Sanders *et al.*, 2000). The WEKA default settings of these techniques are used in this examination (Hall *et al.*, 2009).

Table 1. Summary of Literature identified with cross-validation

S.No	Researcher	Studies	Limitation of Work	Techniques
1.	(Kohavi, R., 1995)	A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.	The results of these show that stratification is a better scheme in terms of bias and variation than regular cross- validation. There is a slight variation on some differences in bootstrap, yet there is a very detailed bias. They recommend using stratified cross-verb multiplication for model selection. Due to the absence of space, we ignore some graphs for Naive-Bayes algorithms when behavior is roughly similar to C 4.5.	C4.5 and a Naive-Bayes algorithm

2.	(Braga-Neto, U.M., and Dougherty, E.R., 2004)	Microarray classification normally has two striking characteristics: (1) classifier design, as well as error estimation, are dependent on amazingly little examples and (2) cross-validation error estimation is utilized in most of the papers.	Cross-validation estimators are particularly problematic in little example settings, ordinarily having higher variance than that of resubstituting or bootstrap estimators.	Linear discriminant analysis (LDA), 3-nearest-neighbor (3NN) and decision trees (CART).
3.	(Schaffer, C., 1993)	Precisely points of view cross-validation as a meta-learning strategy that allows us a chance to select which among an assumed set of learners is to give the best predictive model.	These three strategies (C 4.5, C 4.5 rules, and Backpropagation algorithm) are illustrative of different classification ideal models, yet no effort was made to represent all major classification paradigms or to choose the best algorithm in each.	Decision trees, One for rule sets, and Neural Networks
4.	(Dietterich Thomas G., 1998)	Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.	Be that as it may, much of the time, the amount of data is constrained, as well as they have to utilize all we have as input to our learning algorithms.	k-fold cross-validated t-test, Paired t-test strategies, Cross-validation.
5	(Williams *et* . *al*., 2006)	A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification.	The main limitation in choosing the features was that the calculation within a resource-interrupted IP network device should be actually possible. With time, the amount of data is disrupted, as well as the use of the learning algorithm as input.	Naïve Bayes, C4.5, Bayesian Network and Naïve Bayes Tree algorithms.
6.	(Almeida, *et.al.*, 2017)	Comparison Of Machine Learning Algorithms In Weka	In relation to the accuracy of using different datasets comparatively, they compared only four machine learning processes such as Naive Bayes, Multi-Level Perceptron, and J 48.	Naïve Bayes, Multi-level Perceptron and J48.

Several classification techniques have been introduced in Table 1. In this paper our objective is to measure and test specific classification techniques. Their advantages and disadvantages have been compared between different classification algorithms to present the boundaries.

3. Data Set and Data Collection

This exploration uses object oriented dataset maintenance by CK Metrics. These datasets are chosen for the most part, because various machine learning prototypes have been used to assess object-oriented software maintenance. The ability to remove these results is necessary. All datasets include eight class-level metrics, as there are unique dependent variables with seven independent variables. The independent variable is the seven Chidambar and Kemerar Metrics (Elish and Elish, 2009). In cases, three thousand five hundred eighteen, as well as eight properties, WMC, DIT, NOC, CBO, RFC, LCOM, LOC, dependent variables, a maintenance achievement option are quantities (changes), and the number of lines received in the code we do. They are converted into per square in every class age. One line change may be an extension or else, cancellation or change exists. With that change, the substance of a line is considered when there is an erosion, expansion and change.

Table 2 Characteristics of data-sets.

Dataset	**Total instance**	**Input Classes**	**Output attributes**
5 Open Source Java	3518	7	1

In Table 2, Databases, formats and protocols, games and entertainment, multimedia, science and engineering domain, five open source software written in Java programming. In addition to the total sample of 8 features (7 for input and 1 for output) with 3518, open source has been analyzed through Java medium. Again, according to the purity of prototype and F- Measure , it is evaluated and performed using two different standards (Kohavi, 1995). The internal features of the software are exclusive independent variables used within the LOC with research with specific WMC, DIT, NOC, CBO, RFC, LCOM. The metric starts with various metric suits. Object oriented metrics for the independent variable in the prediction model have been highlighted, available for the initial period of software development. Internal features of open source software framework (metrics) have appeared in additional details in Table 1.

Table 3. Explanation of Metrics suites of open source software frameworks

Metrics Name	Explanation
Weighted Methods per Class (WMC)	WMC strategies the entire complexity of the class. This was the summation of entire complexities of its strategies.
The depth of Inheritance Tree (DIT)	DIT of a class in an inheritance chain of command was the most extreme distance against the class node via the root of the tree. These proposals as long as all class a level from the inheritance stages starting the object chain of
Number Of Children (NOC)	This aggregates the sum of classes and that acquire a specific class. This quantifies the number of urgent descendants of the class.

Coupling Between Object (CBO)	This was clear by way of the aggregate number of different classes to and that class has been coupled. This has been quantifying the number of classes coupled with an accustomed class.
Response For Class (RFC)	Response For Class continues the calculation of the conventional of altogether techniques as would be theoretically exist raised in response facing entire strategies available inside the class chain of command.
Lack of Cohesion in Methods (LCOM)	Lack of Cohesion in Methods do a check about the quantity of techniques sets wherever comparability obtains 0 less the sum of strategy sets anywhere resemblance exists not nil.
Lines of Code (LOC)	It measures the number of lines of code in a class.

In Table 3, Open source programming is programming with source code that anybody can investigate, alter, and improve. "Source code" is the piece of programming that most PC clients absolutely never observe; it's the code PC developers can control to change how a bit of programming —a "program" or "application"—works. This metric suite was proposed by Chidamber & Kemerer in 1994. The suite utilizes the class as a crucial component of object-oriented framework. The long use understanding of the suite has demonstrated its effectiveness and demonstrated that it’s performed well. It’s across the board utilize just affirms it.

4. Classification Metrics

To evaluate the prediction model for achieving the results of the appropriate appraisal, it has been executed 10 times with cross- validation. The similarity between classifications depends on the double measurement, especially depending on the F- Measure. F- Measure is the weighted average of precision and small. Therefore, this score takes both false positives and false negatives into account. If false negativity and false negatives have the same cost, then purity works best. If the cost of false positive and false negatives is very different, then it is better to see both Precision and Remembrance.

F –Measure = 2*(Recall * Precision) / (Recall + Precision)

5. Methodology

The execution of the prototype has been evaluated ten times through cross- validation. Considering this exploration, Chidambar, Kemerer Metrics Suits, as well as LOC Metric, Naive Bayes, ADT, Decision Table, Hyper Pipes are reflected to create prototypes using classified classification processes. Classification, the above mentioned cases have been used before taking prototype. Therefore, tests of classes and identity confusing information can be identified. The classification process consists of a few steps:

- Categorized with set-up and classified class properties of the generated data set
- Classified property properties (importance check).
- In the above set took prototype using the preparation of cases.
- Prototype is used to organize a suspicious data sample.

A model is created by the method of classification, which depicts different sections of data so that the classes are resolved. After classification techniques, special ADT, decision table, hyper pipes, Naive Bayes have been used in this investigation.

Classification Performance Summary

While assessing the machine learning algorithm, the process is executed on large datasets. Classification seems to be the most prevalent type, which is such a large number of different approaches to consider the implementation of classification algorithms. In the summary of the performance of the classification algorithm, three things are worth noting:

❖ **Classification accuracy:** This is the proportion of the number of correct predictions in the form of a prediction, where 100% of the best algorithms can be completed. These are unequal classes, so there is a need to check the Kappa metric, which displays the same data keeping in mind the balance of the class.

❖ **Accuracy of class**: Inspecting false-positive rates for predictions for the actual class as well as for each class, which has been informed about class breakdown for this issue, informs about unequal or more than two classes not there. If the class is more important than prediction, then this result may be able to understand.

❖ **Confusion matrix**: This is a table in which the number of expectations for each category is displayed. Basically there is a number with each class. It is exceptionally valuable for algorithms to achieve the outline of the mistakes.

6. Analysis, Implementation, and Validation

In this experimental assessment, various machine learning techniques have been analyzed. Its aim is to get better results from the open source framework. The evaluation theory relies on the parameters obtained at the place of Kappa Statistic, Mean Absolute Error (MAE), Root Mean

Square Error (RMSE), Relative Absolute Error and Root Relative Error. Twenty classification estimates that the result is meeting the healthy accuracy of the prototype (Hall and Frank, 2008). In addition, open source Java information is being used with three specific specifications (seven for input, then unique to production), with three thousand five hundred eighteen tests. For example, with the accuracy, F- Measure is assessing the model's performance with specific performance standards. The Hyper Pipes Classifier has the framework of Naïve Bayes, ADT, Decision Table and Iris Relationships. There were eight properties such as WMC, DIT, NOC, CBO, RFC, LCOM, LOC, and Change. To evaluate the prediction prototype right before the process, just about ten times the cross- validation is done for the results of the evaluation.

Table 4. Different performance metrics running in WEKA (Accuracy by Class)

Classifier		**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
Naïve Bayes		0.951	0.647	0.933	0.951	0.942	0.773	0
		0.353	0.049	0.433	0.353	0.389	0.773	1
	Weighted	0.894	0.59	**0.885**	0.894	**0.889**	0.773	
ADTree		0.818	0.324	0.831	0.818	0.824	0.835	c0
		0.676	0.182	0.657	0.676	0.667	0.835	c1
	Weighted	0.77	0.275	**0.772**	0.77	**0.771**	0.835	
Decision Table		0.818	0.324	0.831	0.818	0.824	0.83	c0
		0.676	0.182	0.657	0.676	0.667	0.83	c1
	Weighted	0.77	0.275	**0.772**	0.77	**0.771**	0.83	
Hyper Pipes		1	1	0.66	1	0.795	0.5	c0
		0	0	0	0	0	0.5	c1
	Weighted	0.66	0.66	**0.436**	0.66	**0.525**	0.5	

In Table 4, the F Measure that is the correctness of the test and its value is 0.889, 0.771, 0.771 and 0.525, respectively Naive Bayes, ADT, Decision Table and Hyper Pipe. The weighted harmonic mean of precision describes as a reminder of the test. ROC curve is a basic tool for analytical testing appraisal. In the ROC curve, the actual positive rate (sensitivity) is applied to the work of false positive rate (100-specificity) for various cut off purposes of the parameter. Accuracy and precision, accuracy reflects proximity of value for standard or recognized value.

Table 5. Error measurement for different classifiers in WEKA

Classifier	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Kappa statistic
Naïve Bayes	**0.3108**	**0.3921**	**68.9937 %**	**82.6609 %**	**0.4838**
ADTree	0.3219	0.398	71.4569 %	83.9158 %	0.4912
Decision Table	**0.3316**	**0.4002**	**73.6199 %**	**84.3668 %**	**0.4912**
HyperPipes	0.5	0.5	98.0039 %	95.4181 %	0

Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error ,and Root relative squared error respectively, on the assessment of the Naive Bayes classifier in Table 5, respectively, 0.4838, 0.3108, 0.392, 68.9937% and 82.6609 %

Regression performance summary

For the regression problem, we did various performance measures for auditing. There are two things to summarize the performance of the regression algorithm:

❖ **Correlation Coefficient:** This is the way by which the predictions coincide with actual yield value or variation. The value of 0 is most notable, and the value of 1 is an indisputable set of predictions.

❖ **Root Mean Square Error:** This yield is the average amount of error generated on the test set in units of variable.

7. Threats to Validity

In this area, the main hazards facing the validity of this experimental research have been considered and it has been examined.

❖ To reduce the expected validity, a cross-verification test has been done 10 times to achieve continuous results.

❖ Dual measures are used to assess the performance of different categories, such as for accuracy and evaluation of F-measurement.

❖ To reduce the development validity, already approved and major software metrics have been used. The metrics used are acceptable. In the forecast, software changes are widely used.

8. Result

The stratified cross- validation classified Naive Bayes has a dataset of up to 89.369% up to three thousand one hundred and forty four. Misclassified examples are 10.631% of which three hundred seventy-four. These papers show the degree of sample of appropriate and incorrect classified examples. The degree of proper classified specimens is often entitled to accuracy. According to the Naive Bayes Classifier Assessment principles in Table 5, Reviewed the status of Kappa Statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error, and Root relative squared error is 0.4838, 0.3108, 0.3921, 68.9937 % as well as 82.6609 % correspondingly. In Table 4, Naïve Bayes, ADT, Decision Table and Hyper Pipe F-Measure 0.889, 0.771, 0.771 and 0.525respectively. The results show that the Naïve Bayes classified prediction models can meet healthy accuracy. High precision is related to less false positive rate. The decision table is found in Table 4, Naïve Bayes, ADT, Decision Table and Hyper Pipe 0.885, 0.772, 0.772, and 0.436 precision, which is very good.

9. Conclusion and Future work

Data mining, classification algorithms (Naive Bayes, ADT, Decision Trees, and Hyper Pipes) are the results of the procedures outlined on the complexity of the time. The results show that through the execution of different classifications in the best available techniques, an important step to increase the quality of preprocessing mining has been found to predict the expanded accuracy of Naive Bayes classifier model. In any case; on the suitability of classification, we have speculated that WEKA Toolkit is the best tool for the ability to run selected classifier. This investigation can be supplemented by employing the use of genetic algorithms, customized annealing and soft computing strategies.

References

1. Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B. and Mock, S., 2004, June. Kepler: an extensible system for design and execution of scientific workflows. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on* (pp. 423-424). IEEE.

2. Braga-Neto, U.M., and Dougherty, E.R., 2004. Is cross-validation valid for small-sample microarray classification?. *Bioinformatics*, *20*(3), pp.374-380.

3. Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM.

4. Chidamber, S.R. and Kemerer, C.F., 1994. A metrics suite for object oriented

design. *IEEE Transactions on software engineering*, *20*(6), pp.476-493.

5. Dietterich, T.G., Lathrop, R.H. and Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, *89*(1-2), pp.31-71.

6. Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), pp.1895-1923.

7. Dong, L., Frank, E., and Kramer, S., 2005, October. Ensembles of balanced nested dichotomies for multi-class problems. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 84-95). Springer, Berlin, Heidelberg.

8. Elish, M.O. and Elish, K.O., 2009, March. Application of treenet in predicting object-oriented software maintainability: A comparative study. In *Software Maintenance and Reengineering, 2009. CSMR'09. 13th European Conference on* (pp. 69-78). IEEE.

9. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., and Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, *9*(Aug), pp.1871-1874.

10. Goebel, M. and Gruenwald, L., 1999. A survey of data mining and knowledge discovery software tools. *ACM SIGKDD explorations newsletter*, *1*(1), pp.20-33.

11. Hall, M.A., and Frank, E., 2008, May. Combining Naive Bayes and Decision Tables. In *FLAIRS conference* (Vol. 2118, pp. 318-319).

12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), pp.10-18.

13. Hornik, K., Buchta, C., and Zeileis, A., 2009. Open-source machine learning: R meets Weka. *Computational Statistics*, *24*(2), pp.225-232.

14. Khoussainov, R., Zuo, X. and Kushmerick, N., 2004. Grid-enabled Weka: A toolkit for machine learning on the grid. *ERCIM news*, *59*(October).

15. Kaur, U., and Singh, G., 2016. Predicting the Behaviour of Open Source Software using Object-Oriented Metrics. *International Journal of Computer Applications*, *150*(5).

16. Kohavi, R., 1995, August. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*(Vol. 14, No. 2, pp. 1137-1145).

17. Rodriguez, J.D., Perez, A. and Lozano, J.A., 2010. Sensitivity analyses of k-fold cross-validation in prediction error estimation. *IEEE transactions on pattern analysis and machine*

intelligence, *32*(3), pp.569-575.

18. Sanders, G.D., Nease Jr, R.F. and Owens, D.K., 2000. Design and pilot evaluation of a system to develop computer-based site-specific practice guidelines from decision models. *Medical Decision Making*, *20*(2), pp.145-159.

19. Sharma, A.K. and Sahni, S., 2011. A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering*, *3*(5), pp.1890-1895.

20. Schaffer, C., 1993. Selecting a classification method by cross-validation. *Machine Learning*, *13*(1), pp.135-143

21. Stumpf, S., Rajaram, V., Li, L., Wong, W.K., Burnett, M., Dietterich, T., Sullivan, E. and Herlocker, J., 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, *67*(8), pp.639-662.

22. Wahbeh, A.H., Al-Radaideh, Q.A., Al-Kabi, M.N., and Al-Shawakfa, E.M., 2011. A comparison study between data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications*, *8*(2), pp.18-26.

23. Williams, N., Zander, S. and Armitage, G., 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, *36*(5), pp.5-16.

24. Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

25. WEKA, the University of Waikato, Available at http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 April 2011).