JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis of Different Data Mining Models: Decision Tree, Meta and Naïve Bayes

Abhishek, Dr. Dinesh Kumar
M.Tech Scholar, Professor
Department of CSE
BRCM CET, Bahal,(Haryana),India

ABSTRACT

Data mining problems that are concerned with prediction fall under the category of classification. Its primary responsibility is to categories the data to derive predictions about future data. The act or process of classifying anything might be thought of as classification in general. Building models that predict an object's class based on its features is one of the most prevalent tasks in data mining.The classification algorithms Naive Bayes, Meta Classifiers and Decision Tree Based Classifiers have all been compared in this research. To assess the effectiveness of several categorization algorithms, an experiment has been put up. Theoretical study and experimental findings demonstrate that the "Classification by Regression" method successfully identified all examples, while "Decision Tree based Classifiers" produced the fewest errors and "Nave Bayes" classified all instances in the shortest amount of time.

IndexTerms **- Naïve Bayes, Datasets, Classification, Meta Classifiers, Naïve Bayes, Decision Tree.**

INTRODUCTION

Classification is one of the best applications of machine learning algorithms, which applies to the general problem of supervised learning where a given set of training datasets is classified to one or more predefined categories. The main aim of classification is to classify the datasets; even when the class label of the dataset is unknown. This process can be related with the similar one i.e. prediction. Prediction has many applications; such as weather forecasting. In prediction, classification is used to predict the class of a specific instance of a dataset. In some cases, we might combine Clustering and Regression to complete the same objective. Data mining, as we all know, is a crucial phase in the "Knowledge Discovery from Databases (KDD)" method, a relatively new and diverse area of computer science that seeks to unearth intriguing but concealed patterns in enormous data sets. It uses techniques from the study of statistics, database management, machine learning, and artificial intelligence. The ultimate purpose of the data mining is to remove insights from data and convert it into a usable structure. In addition to the raw analysis step, it includes database and data management methods, data analysis, models and inferences factors, complexity concerns, data preprocessing of identified structures, presentation, and online updating. Data mining tasks that are usually applied [1] are classed as follows: Categorization is the task of generalizing well-known framework to use to updated information that does not have a classification. For example, record classification based on the 'class' attribute. Projection and extrapolation are also included in classification techniques. Clustering is the task of discovering groupings commonalities of datasets inside this clustered and incongruities well outside groupings from the data gathering. Clustering approaches also include anomaly detection (outlier/change/deviation detection). This stage is typically used to identify unusual/abnormal data records or errors, which might be entertaining at times. Outliers in any situation may necessitate further study and processing. The task of finding meaningful relationships between multiple properties of a dataset is known as association rule mining (Dependency modelling). Associations are usually focused on newly discovered, exciting yet hidden patterns. A restaurant, for example, could gather data on customer buying patterns. The store can utilise association rule learning to understand which products are usually purchased combined and will use this content to customers. This is also known as the market bundle study. Based on the features of the dataset or the efficacy of other learning algorithms, meta-learning attempts to forecast the optimal method for a given problem.***Naïve Bayes*** Classifiers and Meta-Level Classifiers are two types of classification algorithms in Meta- classification.

There has been a lot of study done in the area of classification and clustering, however the proposed work will examine the effectiveness of Nave Bayes, Decision Tree based Classifiers, and Meta Classifiers (Classification via Clustering, Classification via Regression, Filtered Classifiers). Due to various their versatility, meta classifiers are now often used in real-world settings. Meta classifiers include, for instance, Feature Selection Classifier, Classification via Correlation, Multi - class classification Classifiers, Stochastic Feature Space Classifiers, and Tainted Classifiers. Filtered classifiers can use a variety of filters, both supervised and unsupervised. On the basis of the database, filters can be classified as attribute-based filters or instance-based filters. The recommended effort will solely use feature supervised classifiers to metaclassify datasets. Bayesian classifiers, on either hand, are illustrated by Nave Bayes classifiers. Bayes Net and Nave Bayes are two popular Bayesian classifiers. In the following chapters, we will go through each classifier in depth. As is common knowledge, trees are the best way to classify anything. Of all the classes, trees serve as the best examples. We may prefer to employ the well accepted J48 approach. This paper's effort will investigate the quality of the selected classification methods: Classification via Clustering:

1. Regression Classification
2. Bayesian Classifiers
3. Classifiers based on Decision Trees
4. Classifiers with Filters

The following part provides the preliminary work connected to the different classifiers, while section III explain on the research investigation and chapter Iv discusses the concluding remarks study.

BACKGROUND WORK

Over the years, a number of classification methods have surfaced that explain how to get the best results possible from classification systems. Neural Networks (NNs), Bayesian Networks, Decision Trees (DTs), Naive Bayes (NBs), Support Vector Machines (SVMs), and others are a few of them. Only the three widely used classification methods—NBs, DTs, and Meta Classification—will be discussed in this study. Despite the advantages of various strategies, our research is related to a small business. A Nave Bayes classifier is a straightforward probabilistic classifier that uses Bayes' theorem [4, 9, 13]. The Nave Bayes classifier has the benefit of requiring only a little quantity of learning approach to predict the parameters, i.e. the averages and eccentricities of the variables, required for classification. Due to the assumption of independent variables, just the variances of the variables for each class must be determined. A decision-support tool known as a decision tree utilizes a tree-like graph or depiction of options and their potential scenarios, such as value, overhead charges, and random event outcomes. It is one way to show an algorithm [on Wikipedia]. Some of the work on tree-based classifiers can be found in [7, 12]. We used the concepts of Classification via Clustering, Classification via Regression, and Filtered Classification in Meta Classification.

EXPERIMENTAL WORK

In order to validate the analysis of different classifiers, a number of experiments have been set up. All experiments have used four real-world datasets from the weka [3] repository: Diabetes, Vote, Glass, and Weather. Table 1 contains information on the datasets. All tests were conducted using an Intel(R) Core(TM) i5 CPU @ 2.5 GHz processor and 8 gigabytes RAM. As a development tool for clustering data objects, we used WEKA 3.9.4. During the experimental setup, we have tested aforesaid algorithms of classification on these four datasets. Throughout the entire experiment, five fundamental parameters were recorded and are listed below:

1. Cases that were categorized properly
2. Cases that were incorrectly labelled
3. the Kappa value
4. Mean absolute error, as well as
5. Error in the root mean square

Table 1: Characteristics of various datasets that are used during this work

Relation Name	Number of Instances	Type of Data
Weather	14	Nominal
Vote	435	Nominal
Diabetes	768	Numeric
Glass	214	Numeric

A summary of all the experiments is shown in figure 1. The values provided against each classifier in figure 1are average of all the values obtained from 4 datasets. For example, the value of Classification via Clustering under correctly classified instances is 61.8025, which is the average of all the correctly classified instances of all 4 data sets. Classification via clustering algorithm has classified 57.14 % of all the instances of ‘Weather’ dataset, 85.05 % of ‘Vote’ dataset,40.18 % of ‘Glass’ dataset and 64.84 % of ‘Diabetes’ dataset. The average value of all these values is 61.80 which are mentioned in the figure 1

Average values of 5 parameters on 4 datasets by different algorithms					
	Classification via Clustering	Classification via Regression	Naïve Bayes Classifiers	Decision Tree based Classifiers	Filtered Classifiers
Correctly Classified Instances (in %)	61.4825	76.585	68.695	71.74	75.14
Incorrectly Classified Instances (in %)	38.1675	23.4825	31.955	28.2525	24.85
Kappa Statistic	0.26855	0.537725	0.366425	0.45285	0.5084
Mean Absolute Error	0.275125	0.23545	0.245775	0.22485	0.24235
Root Mean Squared Error	0.5119	0.34885	0.3855	0.3773	0.3513725
Time Taken to build model (in sec)	0.04	0.2625	0.085	0.0175	0.0125

Figure 1: Average values of each parameter by five different classification algorithms on four real datasets

After analyzing all the values of figure 1, we can say that 'Classification via Regression' algorithm has the maximum value for correctly classified instances parameter which is also shown in figure 2. On the other hand, if we talk about the incorrectly classified instances parameter, the same kind of lead has been taken by 'Classification via Regression' algorithm, which is having the lesser value among all the algorithms as shown in figure 3. Also the maximum value of Kappa Statistic has been gained by the 'Classification via Regression' method, which is clearly shown in figure2.

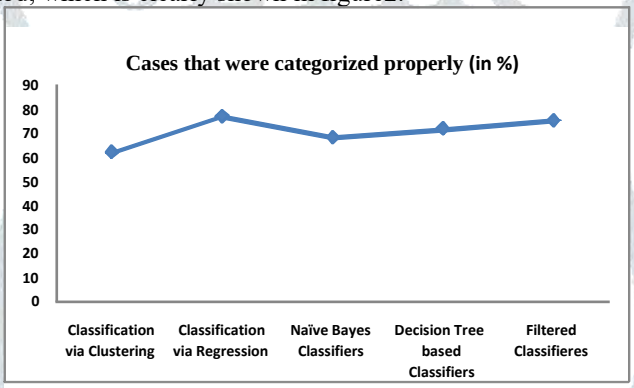

Figure 2: Percentage of Cases that were categorized properly (in %) by all 5 algorithms

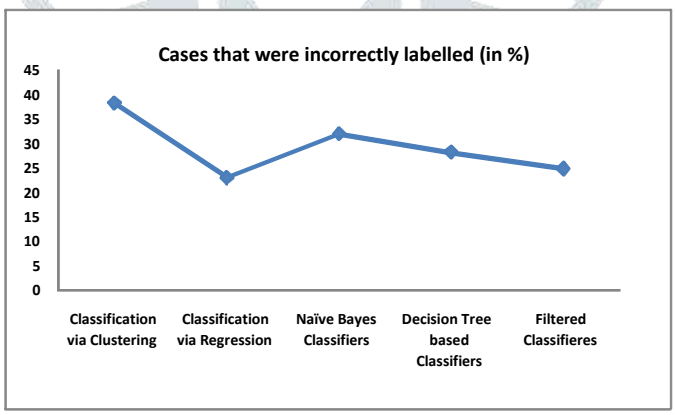

Figure 3: Percentage of Cases that were incorrectly labelled by all 5 algorithms

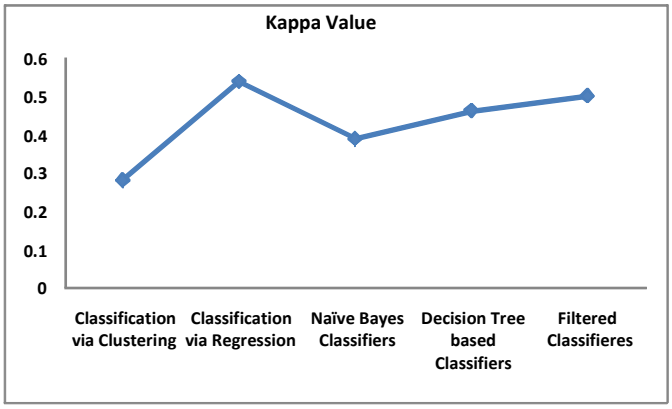

Figure 4: Kappa value of all the 5 algorithms on 4 datasets

As per the errors are concerned the minimum errors has been produced by 'Decision Tree based Classifiers' which can be easily seen in figures 5 and 6, whereas all the classifications have been done in minimum time by 'Naïve Bayes' algorithm. Figure 7 shows the time taken by all the classifiers on the 4 real world datasets.

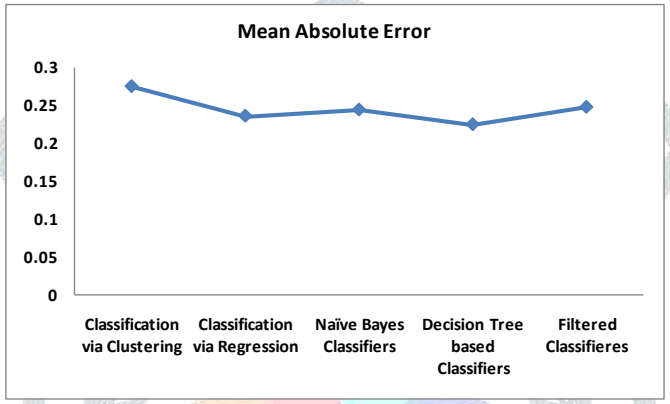

Figure 5: Mean absolute errors generated by all 5 algorithms

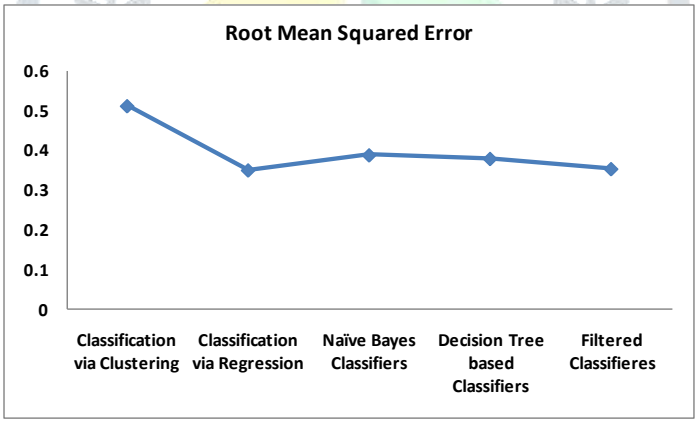

Figure 6: Root mean squared errors generated by all 5 algorithms

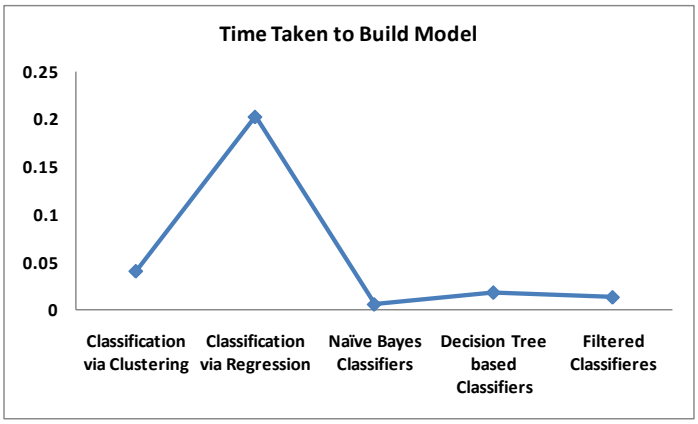

Figure 7: Time taken by all 5 algorithms in classification

CONCLUSIONS

The analysis of five categorization methods using four real-world datasets is the main emphasis of this paper. Theoretical study and experimental findings demonstrate that the "Classification by Regression" method successfully identified all examples, while "Decision Tree based Classifiers" produced the fewest errors and "Nave Bayes" classified all instances in the shortest amount of time. This research could be expanded by taking into account various algorithms with large real-world datasets.

REFERENCES

[1] Aggarwal, Charu C., and ChengXiangZhai. "A survey of text classification algorithms." *Mining text data*. Springer US, 2019.163-222.

[2] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, USA,2018.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann& I. H. Witten (2018); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[4] Muralidharan, V., and V. Sugumaran. "A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis." *Applied Soft Computing* 12.8 (2018): 2023-2029.

[5] Thornton, Chris, et al. "Auto-WEKA: Combined selection and hyperparameter optimizationof classification algorithms." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2017.

[6] Kou, Gang, et al. "Evaluation of classification algorithms using MCDM and rank correlation." International Journal of Information Technology & Decision Making 11.01 (2017): 197-225.

[7] Kotsiantis, Sotiris B. "Decision trees: a recent overview."Artificial Intelligence Review 39.4 (2015): 261-283.

[8] Fernández-Blanco, Enrique, et al. "Random Forest classification based on star graph topological indices for antioxidant proteins." Journal of theoretical biology 317 (2016): 331-337.

[9] Sharma, Sanjay Kumar, et al. "An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification." Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on. IEEE, 2016.

[10] Barnaghi, PeimanMamani, VahidAlizadehSahzabi, and Azuraliza Abu Bakar. "A comparative study for various methods of classification." Int. Conference on Information and Computer Networks (ICICN 2016). IPCSIT. Vol. 27. 2014.

[11] Wang, Xi-Zhao, Ling-Cai Dong, and Jian-Hui Yan. "Maximum ambiguity-based sample selection in fuzzy decision tree induction." Knowledge and Data Engineering, IEEE Transactions on 24.8 (2015): 1491- 1505.

[12] Kalpana, Saravanan and Vivekanandan, "A Two-Stage Tree based Meta-Classifier using Stack-Generalization", International Journal of Computer Applications (0975 – 8887), Volume 36– No.3, December 2014, pp. 25-28.

[13] Daniela, Christopher and Roger, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Web Pages", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2014, pp. 13-23

[14] LiangZhao, Deng-Feng Chen, Sheng-Jun Xu and Jun Lu, "The Research of Data Mining Classification Algorithm that Based on SJEP", International Journal of Database Theory and Application, Vol.8, No.2, pp.223-234, 2019.

[15] Anita Ganpati, "A Performance Comparison Of End, Bagging and Dagging Meta Classification Algorithms", Proceedings of Academics World 24th International Conference, 2018

[16] Ali Selamat Faculty and NurulhudaZainuddin "Sentiment Analysis Using Support Vector Machine" 2019 IEEE 2019 International Conference on Computer, Communication, and Control Technology (I4CT 2019), September 2 - 4, 2019 - Langkawi, Kedah, Malaysia978.

[17] W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0

[18] M. Welling, "A First Encounter with Machine Learning"
M. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118-96174-2.

[19] ChitraNasa, Suman "Evaluation of Different Classification Techniques for WEB Data "International Journal of Computer Applications (0975 – 8887) Volume 52– No.9, August 2012.

[20] Sandhya N. dhage, Sandhya N. dhage "A review on Machine Learning Techniques" I nternational Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

[21] AyonDey "Machine Learning Algorithms: A Review" International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179.

[22] S. B. Kotsiantis "Supervised Machine Learning: A Review of Classification Techniques" Informatica 31(2007) 249-268

[23] V. Sharma, S. Rai, A. Dev, "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277128X, Volume 2, Issue 10, October 2012.