Significance of Attributes to Explore Accuracy Using Various Classifiers

Shraddha Kavathekar
ME (Computer Engineering)1st year
PCCOE, akurdi, Pune, India

Shraddha Tadmare
ME (Computer Engineering)1st year
PCCOE, akurdi, Pune, India

Dhanshri Kamble
ME (Computer Engineering)1st year
PCCOE, akurdi, Pune, India

ABSTRACT

Data mining is a process of extracting knowledge from a huge set of data. The major components of data mining techniques are classification, association rules and sequential analysis. Classification is an important data mining technique with immense applications to classify the different kinds of data used in nearly every field of our life. Classification is a data mining (machine learning) technique used to identify group membership for data occasion. In this paper the basic classification techniques such as Naïve Bayes, Naïve_bayes_simple, Random Forest, and J48 are experimented. The aim of this study is to provide an interdisciplinary review of different classification techniques in data mining and use of variety of datasets such as iris, diabetes and soybean for experimenting accuracy of divergent classifiers. Also, the paper discusses on selection of attributes based on their significance using best first search method.

Keywords
Data mining, Classification, j48, Naïve Bayes, Naïve_Bayes_simple, Random forest classifier, Attribute selection, Accuracy.

I INTRODUCTION

1.1 Data Mining

Data mining [3] is an approach to perceive interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from tremendous amount of data stored in information repositories. Data mining is a major elevation in the type of analytical tools. Data mining is a multi-disciplinary field which is an integration of machine learning, statistics, database technology and artificial intelligence. This technique includes number of phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. There are 5 data mining techniques such as Association, Classification, Clustering, Neural Network and Regression.

1.2 Classification[5]

Classification is used to stratify the item according to the features of the item with respect to the predefined set of classes. Classification is a data mining (machine learning) technique used to envisage group membership for data instances. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be enrooted based on scrutinize data for many loan applicants over a period of time. This paper compares the different classifiers with their accuracy for variant datasets.

1.3 Algorithms

1.3.1 Naïve Bayes

The Naive Bayes algorithm [5] is a simple contingency classifier that calculates a set of probabilities by counting the frequency and consolidation of values in a given data set. The algorithm uses Bayes theorem and estimate all attributes to be independent given the value of the class variable. The simulation as Naive yet the algorithm tends to perform well and learn speedily in various supervised classification. It performs different applications such as sentiment analysis, document categorization and email span filtering Naive Bayesian classifier is deployed on Bayes theorem and the theorem of total probability.

$$P(C|X) = \frac{P(X|C).\,P(C)}{P(X)} \quad \ldots\ldots (1)$$

Where P (C|X) is the posterior probability, P(C) is the prior probability, P(X) is predictor prior probability [2]

1.3.2 J48 [1]

J48 classifier is a simple C4.5 decision tree for stratification. In the Weka tool, it is an open source java implementation of C4.5 algorithm. With this technique, a binary tree is composed to model the classification process. Once the tree is created, it is applied to each tuple in the database and results in categorization for that tuple. The supplementary features of J48 are handling missing values, decision trees pruning, continuous attribute value ranges, extraction of rules, etc.

1.3.3 Random Forest [8]

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean forecasting (regression) of the individual trees. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. It runs efficiently on large datasets and handles thousands of input variables without variable deletion or any errors. It gives estimates of what variables are important in the classification. It uses a Bagging approach to create a bunch of decision trees with random subset of data. The output of decision tress in the random forests is combined to make the final prediction. the final of the random forest algorithm is extracted - by surveying the results of each decision trees and just by going with prediction that appears the most times in decision trees.

II PROPOSED METHOD

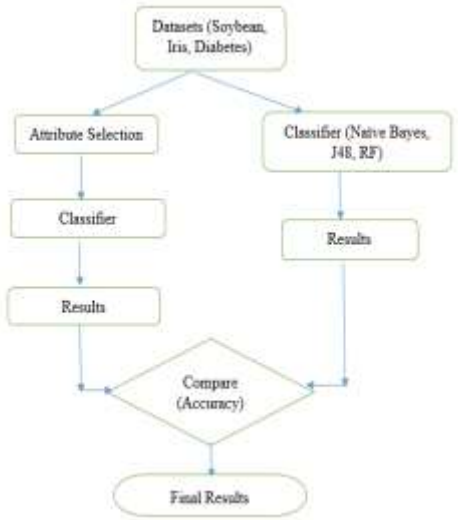


Figure 1. Working Model

Fig (1) shows Different datasets are selected for classification such as iris, diabetes and soybean. Classification techniques used are naïve Bayes, j48, random forests. This paper estimates the accuracy for the above three datasets and comparison is done for different classifiers. The attributes are selected and their accuracy is calculated as shown in the below tables to see whether the accuracy is reduced if the attributes are reduced and which classifier gives highest accuracy for the three datasets. Some of the attributes are pruned from the datasets and their accuracy is calculated. Comparison is done to see whether the classifier maintains the same accuracy without pruning selected attributes.

II DATA SETS

We have used different datasets and tested in Weka to check their accuracy for different classifiers. Datasets were taken from UCI Machine Learning Repository .

1.1 Iris Dataset

The Iris dataset contains different types of irises (Sentosa, versicolour, verginika) petal and sepal length stored in 150*4 numpy, ndarray. This is feasibly the best-known database to be found in the pattern recognition works. In the iris dataset the Number of attributes are 4, Number of instances are 150 and the Attribute characteristic of whether dataset is real. [17]

1.2 Soybean Dataset

In the soybean dataset the Number of attributes are 35, Number of instances are 47 and the Attribute characteristic of whether dataset is categorical. [18]

1.3 Diabetes

In the diabetes dataset the Attributes are 9, including instances 768. The different attributes of dataset are preg, plas, pres, skin, class[19]

III EXPERIMENTAL RESULTS

This paper presents different classifiers in weka and testing different datasets for above algorithms. Following table described their accuracy. The result shows that Iris dataset sworks well with Naïve Bayes as compared to Naive_Bayes_Simple, Random forests and J48. For soybean dataset accuracy is calculated for the entire dataset and by selecting random attributes the results show that for soybean dataset when some random attributes pruned then accuracy is reduced. Naive bayes has maximum accuracy 92.97% for soybean dataset. For iris dataset the results are similar for whole dataset as well as for selected attributes. Naive bayes gives highest accuracy 96%. For diabetes dataset when attributes are randomly selected their accuracy is reduced, whereas for whole dataset it is increased and for naïve bayes is shows highest accuracy 76.30%. The overall observation of this paper gives, Accuracy for Naïve Bayes is highest 96% in comparison with other classifiers as showed in below table. The highest accuracy for datasets for varied classifiers in visualized in the table.

The formula to calculate accuracy is:

1. $$\text{Accuracy} = \frac{TP+TN}{P+N} \quad \text{----- (2)}$$

In the equations (2) above Accuracy represents Total Accuracy, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Table 1.1. Operations on Soybean Dataset

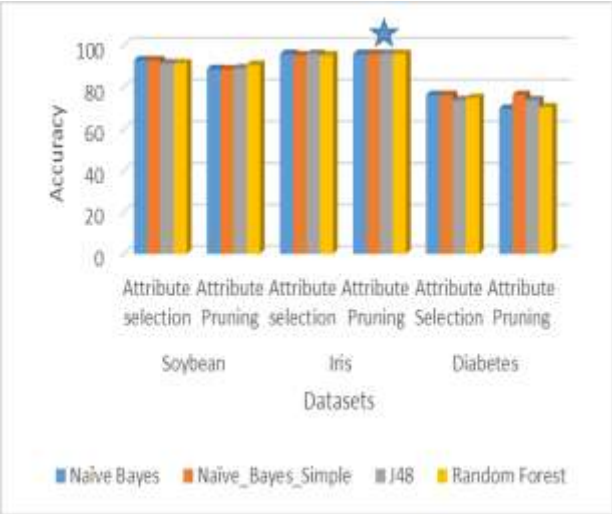
Algorithm Used	Attribute Selection	Attribute Pruning
Naïve Bayes	92.97	88.72
Naïve Bayes simple	92.97	88.57
J48	91.50	89.16
Random Forest	91.50	90.77

Table1.2. Operations on Iris Dataset

Algorithm Used	Attribute Selection	Attribute Pruning
Naïve Bayes	96	96
Naïve Bayes simple	95.33	96
J48	96	96
Random Forest	95.33	96

Table1.3. Operations on Diabetes Dataset

Algorithm Used	Attribute Selection	Attribute Pruning
Naïve Bayes	76.30	69.79
Naïve Bayes simple	76.30	76.30
J48	73.82	73.82
Random Forest	74.86	70.44



IV CONCLUSION

The results obtained demonstrates that the classification performance of four different data mining models - Naive Bayes, Naïve bayes simple, J48, Random Forest are different for varied datasets. When some attributes are pruned randomly the accuracy is reduced. Experimental results have shown the effectiveness of models. Naïve Bayes has highest accuracy for Iris dataset 95.53%. We also calculated error rate for different classifiers for iris dataset. We discovered that the two parameters accuracy and error rate for iris dataset gave different results for different classifiers. Highest accuracy was given by naïve_bayes and lowest error rate was given by J48 4.95.

V REFERENCES

1. Gaganjot Kaur Amit Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes”, International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014
2. Guo, Yang, Guohua Bai, and Yan Hu. “Using Bayes Network for Prediction of Type-2 Diabetes.” In Internet Technology and Secured Transactions, 2012 International Conference For, pp. 471-472. IEEE, 2012.
3. Hany A. Elsalamony, Helwan University, Cairo, “Bank Direct Marketing Analysis of Data Mining Techniques”, Saudi Arabia International Journal of Computer Applications (0975 – 8887Volume 85 – No 7, January 2014
4. A. Floares., A. Birlutiu. “Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis”. WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.
5. Tina R. Patil, Mrs S. S. Sherekar “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification”, International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.
6. Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. Devi Prasad Bhukya1 and S. Ramachandram2.
7. Milos Ilic, Petar Spalevic and Mladen Veinovic, Wejdan Saed Alatresh, “Students’ success prediction using Weka tool”, in INFOTEH-JAHORINA Vol. 15, March 2016.
8. Sharma, Aman Kumar, and Suruchi Sahni. "A comparative study of classification algorithms for spam email data analysis." *International Journal on Computer Science and Engineering* 3.5 (2011): 1890-1895
9. Li, Xiangyang, and Nong Ye. "A supervised clustering and classification algorithm for mining data with mixed variables." *IEEE Transactions on* Systems, Man, and Cybernetics-Part *A:* Systems and Humans 36.2 (2006): 396-406.
10. Yao, Zheng, et al. "R-C4. 5 Decision tree model and its applications to health care dataset." Services Systems and Services Management, *2005.* Proceedings *of* ICSSSM'05. *2005* International Conference *on*. Vol. 2. IEEE, 2005.
11. Tu, My Chau, Dongil Shin, and Dongkyoo Shin. "A comparative study of medical data classification methods based on decision tree and bagging algorithms." Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on. IEEE, 2009.
12. Aher, Sunita B., and L. M. R. J. Lobo. "Data mining in educational system using Weka." *IJCA* Proceedings on International Conference *on Emerging Technology Trends (ICETT)*. Vol. 3. 2011.

13. Umamaheswari, K., and S. Niraimathi. "A study on student data analysis using data mining techniques." *International Journal of Advanced Research in Computer Science and Software Engineering* 3.8 (2013): 117-20.
14. Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." *arXiv preprint arXiv:1206.1121* (2012).
15. Thepade, Sudeep D., and Madhura M. Kalbhor. "Novel data mining based image classification with Bayes, tree, rule, lazy and function classifiers using fractional row mean of cosine, sine and walsh column transformed images." Communication, Information & Computing Technology *(*ICCICT*), 2015* International Conference on. IEEE, 2015.
16. Kasat, Neha R., and Sudeep D. Thepade. "Novel Content Based Image Classification Method Using LBG Vector Quantization Method with Bayes and Lazy Family Data Mining Classifiers." Procedia Computer Science 79 (2016): 483-489.
17. https://archive.ics.uci.edu>suppory>iris
18. https://archive.ics.uci.edu
19. https://archive.ics.uci.edu