

Impact of Classification Algorithms on Census Dataset



Sangavi N, Jeevitha R, Kathirvel P, Premalatha

Abstract: Data mining is a method by which valuable information can be obtained from large databases. A supervised method of classification assigns data samples to target groups. In this system, it uses various classification algorithms namely decision trees, SVM, random forest and neural network. This system will classify and analyses the best suited algorithm which gives maximum accuracy among the other algorithms. The accuracy in these algorithms has been calculated by sensitivity and specificity. Evaluation of these models has been calculated by the error rate with respect to the classes. It uses census dataset and finds whether the income above 50k or below 50k. Matrix of error consists of true positive, neutral, true negative and false negative values. Based on true positive and false negative values, specificity is determined. Based on true negative and false positive values, sensitivity is determined. The algorithm analysis which finds the better algorithm with respect to the accuracy, error rate and efficiency.

Keywords: Decision tree, Neural network model Random forest model, SVM.

I. INTRODUCTION

The dataset of the census includes different attributes such as age, place of job, gender, final weight, marital status, occupation, marriage, sex, capital gain and capital loss. The various models has been used are SVM, random forest, decision tree and neural network. Random forests uses supervised learning method used for classification. Random forest model allows to construction of many decision trees with respect to the variables. Decision tree is used to solve the problem using both regression and classification. Each internal node is an attribute and a class label for each leaf node. SVM is a supervised learning process that uses the hyper plane to sort two groups. Neural network is a thousands of nodes are interconnected used for sales forecasting and time series prediction. There are three layers such as hidden layer, input layer and output layer [8].

II. OBJECTIVE

The project's main objective is to find the performance efficiency and accuracy of the algorithms and to figure out which ones are best suited among those algorithms. The prediction over these algorithms has been calculated with respect to specificity, sensitivity and accuracy. The various models has been considered such as random forest, decision tree, SVM and neural network.

III. PERFORMANCE MEASURE

i) Accuracy

Accuracy is defined by the total number of evaluations as the number of correct evaluations. The specificity and sensitivity values are used to quantify it [1].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is True Positive, TN is True Negative and FN is False Negative.

ii) Sensitivity

It is determined by sum of all negative evaluations from the number of true positive values. The proposition of positive values separated by the no of observation observes tolerance. True positive and false negatives calculate it [1].

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP is True Positive, FN is False Negative

iii) Specificity

Specificity by the maximum number of negative assessments is determined from the number of true negative asseses. Specificity from true negative and false positive values[1] is determined.

$$Specificity = \frac{TN}{TN + FP}$$

where TN is True Negative and FP is False Negative.

IV. METHODOLOGIES

The dataset of the census was chosen for data analysis. It is taken from the database of UCI machine learning [7]. The census dataset's attributes are age, class of job, occupation, capital gain, loss of capital, employment, marital status, gender, sex, marriage, number of school, hours of work and country of origin. Age, capital gain, loss of capital, working hours are continually valued attributes, while age, occupation, employment, marital status, race, sexuality, relationship are attributes that are categorically valued. The classes of census dataset is above 50k or below 50k [2]. There are 14 attributes in the census dataset. Dataset has been executed in dataset which can be target variable, risk variable, ident variable and it can be ignored.

Manuscript published on January 30, 2020.

* Correspondence Author

N. Sangavi*, Pursuing P.G., Department of Computer Science and Engineering, Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, Tamil Nadu, India.

R. Jeevitha, Pursuing P.G., Department of Computer Science and Engineering, Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, Tamil Nadu, India.

P. Kathirvel, Pursuing P.G., Department of Computer Science and Engineering, Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, Tamil Nadu, India.

Dr. K.Premalatha, Professor and Head, Department of Computer Science and Engineering, Bannari Amman Institute of Technology (Autonomous), Sathyamangalam, Erode, Tamil Nadu, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Ignore is chosen when that attribute is not needed to execute the models. If dataset contains more categorical variables then random forest model will not be able to support the model. There are about 32563 records in this dataset.

4.1 DECISION TREE

Decision tree model allows to classify the root nodes and leaf nodes. Root node will classify the nodes and allow to target the nodes.

The leaf node are internal nodes which are attributes of the dataset. The root nodes are the classes which target either 0 or 1. The conditional probability will be checked with respect to the nodes. This model gives the tree representation with respect to the attributes. It is used to represent the root nodes as final classes and other internal nodes are attributes of classes. The relative error and standard error are noted and it will identify the root node. It allows to identify the root node and root node error. Rattle allows to execute the model and time taken to execute the model is noted. Relationship is considered as root node in this dataset[3].

4.2 RANDOM FOREST

Random algorithm forest is a mixture of multiple decision trees. It is a supervised learning technique that allows a maximum number of decision trees to create. The implementation over the decision tree is used in random forests. The decision tree implementation is simple when compared to the random forest. It builds the model and indicates the error and time taken to execute the model. It will not allow categorical variables more than 32 levels. In our dataset, the attribute country has more than 32 categorical values [3].

4.3 SVM

SVM is linear regression used to classify the classes using hyper plane. This model uses line or hyper plane for classification of two sets using training dataset. It is a supervised learning technique that separate the two classes by the hyper plane. It considers some parameters for tuning. They are kernel, gamma, regularization and margin. Misclassification can be avoided by the tuning parameters in SVM. Smaller margin will be optimized easily by the regularization. Kernel allows to solve the equation by the support vector machine. The error of the model is given as relative error. This also allows to know the time taken to complete the execution the model [5].

4.4 NEURAL NETWORK MODEL

Neural Network architecture is a three-layer model. Information is initially given to the layer of information, then to the hidden layer, then to the layer of output. From the secret layer is drawn output layer. It is a bit complex when comparing with other models. Neural network model is derived from biology i.e neurons in our brain. It process the information in parallel and is done with how neurons will work in brain. It is easy for humans to process information in brain. But it is difficult to formulate the information in the brain. One application of neural network is optical character recognition to object detection. It allows the brain that transform the information [5].

V. EVALUATION

5.1 ERROR MATRIX CALCULATION

It should calculate the accuracy, specificity and sensitivity of the models by error matrix

i) Decision tree:

Decision tree's total error rate is 16.3% and the average class error rate is 27.5%

```

Error matrix for the Decision Tree model on adult1.csv [test] (counts):

      Predicted
Actual <=50K >50K Error
<=50K  3502  217  5.8
>50K   579  587  49.7

Error matrix for the Decision Tree model on adult1.csv [test] (proportions)

      Predicted
Actual <=50K >50K Error
<=50K  71.7  4.4  5.8
>50K   11.9 12.0  49.7

Overall error: 16.3%, Averaged class error: 27.75%

Rattle timestamp: 2019-09-25 14:47:12 SANGU
    
```

Figure.1 decision tree error matrix

ii) Random forest

The overall rate of random forest error is 14.1% and the average rate of error in the class is 21.5%

```

Error matrix for the Random Forest model on adult1.csv [validate] (colour)
    Predicted
Actual  <=50K >50K Error
<=50K  3227  220   6.4
>50K   427  740  36.6

Error matrix for the Random Forest model on adult1.csv [validate] (proportions):
    Predicted
Actual  <=50K >50K Error
<=50K  69.9  4.8   6.4
>50K   9.3 16.0  36.6

Overall error: 14.1%, Averaged class error: 21.5%

Rattle timestamp: 2019-10-23 08:55:45 SANGU
=====
    
```

Figure. 2 Random forest error matrix

iii) SVM SVM total error rate is 15.7% and the average class error rate is 24.9%

```

Error matrix for the SVM model on adult1.csv [test] (counts):
    Predicted
Actual  <=50K >50K Error
<=50K  3160  227   6.7
>50K   482  634  43.2

Error matrix for the SVM model on adult1.csv [test] (proportions):
    Predicted
Actual  <=50K >50K Error
<=50K  70.2  5.0   6.7
>50K   10.7 14.1  43.2

Overall error: 15.7%, Averaged class error: 24.95%
    
```

Figure. 3 SVM error matrix

iv) Neural network model The overall error rate for the neural net is 22.8% and the average error rate for the class is 45.65%

```

Error matrix for the Neural Net model on adult1.csv [test] (counts):
    Predicted
Actual  <=50K >50K Error
<=50K  3379   8   0.2
>50K  1017  99  91.1

Error matrix for the Neural Net model on adult1.csv [test] (proportions):
    Predicted
Actual  <=50K >50K Error
<=50K  75.0  0.2   0.2
>50K   22.6  2.2  91.1

Overall error: 22.8%, Averaged class error: 45.65%
    
```

Figure. 4 Neural network error matrix

5.2 SENSITIVITY AND SPECIFICITY

Sensitivity is observed by the proposition of positive values divided by the no of observation. It is calculated by true positive and false negative.

$$Sensitivity = \frac{No\ of\ true\ positive\ assesments}{No\ of\ all\ positive\ assesments}$$

Specificity is calculated from true negative and false positive. It is predictive values of the system compared to the reference results.

$$Specificity = \frac{No\ of\ true\ negative\ assesments}{No\ of\ all\ negative\ assesments}$$

5.3 ACCURACY

Accuracy is calculated by the specificity and sensitivity values. The accuracy among those algorithms has been calculated by error matrix or confusion matrix. True negative, false positive and false negative values are given by the error matrix.

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is True Positive, TN is True Negative and FN is False Negative

Table 5.1 Specificity, Sensitivity and Accuracy Calculation

MODELS	TRUE POSITIVE	FALSE POSITIVE	FALSE NEGATIVE	TRUE NEGATIVE	SENSITIVITY	SPECIFICITY	ACCURACY
Decision tree	3473	191	613	607	85.00	76.07	83.54
random forest	3228	228	442	695	87.96	75.30	85.41
SVM	3230	226	491	646	86.80	74.08	84.39
neural net	3444	12	1022	115	77.12	90.55	77.49

VI. RESULT

In terms of accuracy, specificity and sensitivity the comparison is made with the different models such as decision tree, random forest, SVM and neural network. The accuracy and sensitivity of random forest is higher when comparing other three algorithms. It need not be same for all the datasets. The specificity of neural net is higher than other models. The values which are shown red in color are the higher accuracy, specificity and sensitivity values. The true positive, true negative, false positive and false negative are the values drawn from the error matrix which is used to calculate the accuracy, specificity and sensitivity

Table 6.1 Comparison of various models with respect to accuracy, specificity and sensitivity

Models/ Performance	Accuracy	specificity	Sensitivity
---------------------	----------	-------------	-------------

measure			
Decision tree	83.54	76.07	85.00
Random forest	85.41	75.30	87.96
Svm	84.39	74.08	86.80
Neural network	77.49	90.55	77.12

VII. CONCLUSIONS

This compares different models such as decision tree, random forest, SVM and neural models in this project in order to achieve better accuracy and efficiency results. The analysis of the different models shows the accuracy, specificity and sensitivity performance. The accuracy of the error rate, sensitivity and specificity are measured. Error rate is calculated by a confusion matrix with respect to the values. The accuracy and sensitivity of random forest is higher when comparing other three algorithms. The specificity is higher in neural network model than other algorithms. In future work, the sensitive features are identified and the selected features are preserved by altering their original values with some statistical methods and the performances are analysed with state-of-the art methods.

REFERENCES

1. Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA", International Conference in "Emerging Trends in Science, Technology and Management. 2013;10".
2. A comparative analysis of classification algorithms in datamining for accuracy, speed and robustness "Dogan, N. Technol M (2013) 14:105".
3. S. Archana1, Dr. K. Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications.2014;2(2):11.
4. Dr. A. Bharathi, E. Deepan kumar ,” Survey on Classification Techniques in Data Mining”, International Journal on Recent and Innovation Trends in Computing and Communication. 2014;2(7)
5. Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study,” Decision Support Systems (Elsevier). 2004;37:543– 558. "A Comparative Study of Classification Techniques On Adult Data Set" S.Deepajothi , Dr.S.Selvarajan Chettinad college of Engineering and Technology ,TamilNadu,India
6. "UCI Repository of Machine Learning Databases" by D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, Available at www.ics.uci.edu/~learn/MLRepository.html, University of California, Irvine, 1998.
7. Jiawei Han "Datamining: Concepts and Techniques" Second edition, Morgan

AUTHORS PROFILE



N. Sangavi, currently pursuing Post Graduation in Department of Computer Science and Engineering in Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, TamilNadu. She received B.E degree in SNS College of Technology(Autonomous), Coimbatore. She published 5 research papers in

International Conferences. Her area of interest is data mining and data analytics.





R. Jeevitha, currently pursuing Post Graduation in Department of Computer Science and Engineering in Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, Tamil Nadu. She received B.E degree in Avinashilingam Deemed University, Coimbatore.

She published 5 research papers in International Conferences. Her area of interest is data mining and image processing



P. Kathirvel, currently pursuing Post Graduation in Department of Computer Science and Engineering in Bannari amman Insititute of Technology(Autonomous), Sathyamangalam, Erode, TamilNadu. He received B.E degree, Bannari amman Insititute of Technology(Autonomous), Coimbatore. He published 5 research papers in

International Conferences. His area of interest is data mining and Tensor flow .



Dr. K.Premalatha, currently Professor and Head of Computer Science and Engineering, Bannari Amman Institute of Technology (Autonomous), Sathyamangalam, Erode, TamilNadu. She received her B.E in IRTT Institute and M.E in Kongu Engineering College, Erode. She had nearly 20

years of experience. She served and held many academic positions and as a PG coordinator in Bannari Amman Institute of Technology. She nearly guided 40+ Ph.D students and doing some funded projects in DRDO and DST- NRDMS projects. She has been effectively published 96 journals in IEEE, Springler and Inderscience publishers and 80 Conference papers. Her area of interest is data mining, machine learning and data analytics.