International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-7, Issue-4, April 2017

Comparative Analysis of KNN and C5.0 Algorithm for Smart City Classification

Pragya Sharma, Deep Kumar

Abstract— **Many governments are adopting smart city concept so as to improve living standards. Big data analytics is one of the main technologies that are able to enhance smart city services. As almost everything is becoming digitalized, a lot of data is being collected that can be beneficial in various domains. Various classifications algorithms have been developed in the last decade. Many of them have been compared. This paper presents a comparison between two classification algorithms: KNN and decision tree (C5.0). This paper is aimed to compare two algorithm's results. The results of algorithms are central task in areas such as machine learning. For analyzing the result obtained when comparing the algorithms, the best tool used is RStudio which provides a platform for loading data to produce plots and tables.**

Index Terms— **Classification, Analytics, RStudio, KNN, C5.0**

I. INTRODUCTION

The main strength of the big data concept is the high influence it will have on numerous aspects of a smart city and consequently on people's lives [1]. Smart city adoption is one of the major projects of government. Implementation of big data applications to this project will support various smart city components and will improve living standards. Utilizing various technologies by smart cities will help to improve the performance of education, water services, transportation, medical facilities, electricity and power supply, paved approach roads etc. leading to higher levels of comfort for their citizens. Big Data Analysis provides the ability of handling the data that is obtained from different types of resources to provide quality information. It plays a vital role in census data to classify more accurate results. The research is aimed at utilizing the census data in 2001 to classify weather a city should be under smart city or not. Population data plays an important role in various fields like abortion ratio, electricity supply, water supply, farming, road construction, school development, etc. One of such field is classifying the state as smart city on the basis of various attributes such as population of male and female, their working status, their literacy rate, etc.

Classification is mostly used in the research field. It is best suited for decision-theoretic approaches for predicting data. A data is generally represented by a vector $(x_1,x_2,...,x_n)$ where n is the number of features. So, each vector can be considered as one data. Two essential steps in classification algorithm are Training and Testing. Training data will help to predict the class label of testing data using characteristic properties of training data that is computed by machine learning. There are various types of algorithms that provide the ability of classification of a dataset. But it is very difficult to identify which is the best algorithm. We can only find a conclusion when one classification algorithm surpasses other. This paper presents the experimental analysis of the well-known two classification algorithms: KNN and C5.0 on population dataset. Then the results of both the algorithms have been compared.

The paper is organized as follows: Section 1 provides an introduction to the topic of the research. Section 2 describes tool used in this paper. Section 3 explains the related work studied for this topic. Section 4 gives the overview of the methods that are being used in this paper. Section 5 provides with the experiments and results and finally the conclusion has been conveyed along with the future work.

Pragya Sharma, Department of Computer Science and Engineering, DIT University, Dehradun, Uttarakhand, INDIA

Deep Kumar, Department of Computer Science and Engineering, DIT University, Dehradun, Uttarakhand, INDIA

II. RELATED WORK

It has been said that we have entered the age of Big Data [3]. Only in two years 90% of world's data that is being digitized was captured. As a result, many governments have started to utilize big data to support the development and sustainability of smart cities around the world [4]. There are various smart city characteristics such as city facilities that allow cities to maintain standards, principles, and requirements of the applications of smart city.

In [5] comparison of ten supervised learning algorithms was done. The results were compared using eight performance criteria. They evaluated the performance of many classification problems using variety of performance metrics such as accuracy, squared error, cross-entropy etc. They came to a conclusion that calibrated boosted trees were the best learning algorithm overall.

Another similar approach was done in another paper [6].They compared and analyzed the performance of three machine learning algorithms. This was done to classify human facial expression. The input for this classification process had 23 variables that were calculated from distance of facial features. As a result the output was categorized in seven categories such as happy, neutral, angry, sad, disgust, surprise and fear. They performed some test cases and came to a conclusion that by using smallest amount of data the accuracy was 75.15% for K-Nearest Neighbor (KNN), 80% for Support Vector Machine (SVM), 76.97% for Random Forests algorithm and by using largest amount of data the accuracy was 98.85% for KNN, 90% for SVM, and 98.85% for Random Forests algorithm.

It was demonstrated in [7] that machine learning algorithm can be used to compare the algorithms. They have discussed that machine learning techniques have been used for the classification so as to predict the disease Dengue. They have used two algorithms: SVM, Naïve Bayes. They have discussed the application of machine learning techniques so that Dengue and other diseases can be distinguished like

feverish illness and predict arbovirus among people. They came to a result that SVM outperforms the Naive Bayes in Dengue disease diagnosis.

III. TOOL USED

The analysis of these algorithms is done using the software RStudio that is a free and open-source integrated development environment (IDE) for R and R is a programming language for statistical computing and graphics. R has been ranked as number one tool in Rexer's Survey [2]. RStudio provides various packages that can be installed easily. In this paper class package is being used for KNN algorithm and C50 package for C5.0 algorithm. RStudio is easy to use. It provides auto-completion even as you type R commands, showing various options you can use for the commands.

IV. OVERVIEW OF THE METHODS

Data classification is important in predictive analytics [8][9] and high demanding research area. There are various classifications algorithms such as KNN and decision tree (C5.0).

The most popular algorithm in clasification is KNN. It is found to be very efficient in experiments on datasets. Learning-by-analogy principles are used in Nearest Neighbor classifier. A dataset contains data samples which are to be described by n dimensional numerical attributes. For a given unknown data sample, K- Nearest Neighbor classifier searches n-dimensional space that are closest to the unknown sample by finding its k-Nearest Neighbors with an Euclidian distance measures or Absolute distance measure [10]. Euclidean distance is calculated by the following formula, where p and q are the examples that are to be compared, each having n features. The term p1 is the value of the first feature of example p, while q1 is the value of the first feature of example q

$$\text{dist}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

C5.0 algorithm is a decision tree algorithm that is improved on C4.5 algorithm. In decision Tree Induction, the analysis ability of the tree is stronger when the tree size is smaller. It takes less training time to construct the decision tree, and generated decision tree is interpreted easily.

V. DATASET DISCRIPTION

Census dataset was obtained from the website of Office of the Registrar General & Census Commissioner, India [11] (modified by Kaggle [12]). Almost all the features of this dataset are numeric. Important attributes were kept and rest all were removed. We came up with 38 attributes that could be used for categorizing that a city should be under smart city label or it should be under waiting label [13]. These attributes are as follows:

1. State (Character value)
2. District (Character value)
3. Persons (numeric value)
4. Males (numeric value)
5. Females (numeric value)
6. Growth 1991 to 2001 (numeric value)
7. Number of households (numeric value)
8. Sex ratio females per 1000 males (numeric value)
9. Sex ratio 0-6 years (numeric value)
10. Persons literate (numeric value)
11. Males literate (numeric value)
12. Females (numeric value)
13. Persons literacy rate (numeric value)
14. Males literacy rate (numeric value)
15. Females literacy rate (numeric value)
16. Total educated (numeric value)
17. Matric higher secondary diploma (numeric value)
18. Graduate and above (numeric value)
19. Total workers (numeric value)
20. Main workers (numeric value)
21. Marginal workers (numeric value)
22. Non workers (numeric value)
23. Total inhabited villages (numeric value)
24. Drinking water facilities (numeric value)
25. Safe drinking water (numeric value)
26. Electricity power supply (numeric value)
27. Primary school (numeric value)
28. Middle school (numeric value)
29. Secondary sr. schools (numeric value)
30. Medical facility (numeric value)
31. Primary health Centre (numeric value)
32. Post telegraph and telephone facility (numeric value)
33. Bus services (numeric value)
34. Paved approach road (numeric value)
35. Mud approach road (numeric value)
36. Permanent house (numeric value)
37. Temporary house (numeric value)
38. City label (Character value)

VI. EXPERIMENTS AND RESULTS

In our experiment two algorithms i.e. KNN, C5.0 were implemented on the population dataset. The dataset was divided as train and test data with probability of 0.67 and 0.33 respectively. So, train dataset contains 396 tuples and test dataset contains 194 tuples.

For KNN algorithm, the value of k taken here is 19, an odd number roughly equal to the square root of 396 i.e. number of instances in training data.

A model is created using C5.0 algorithm that contains C5.0 decision tree with size of 14. Number of samples were 396 and number of predictors were 36.

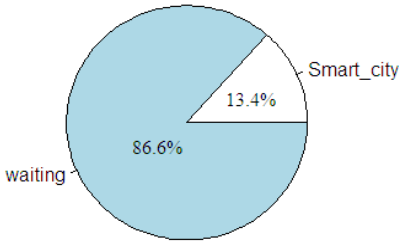
A. *KNN Algorithm*

This Cross Table for KNN algorithm shows that a total of 189 of 194 predictions were true positive. Pie chats were created of both the actual smart cities and waiting cities, predicted smart cities and waiting cities.

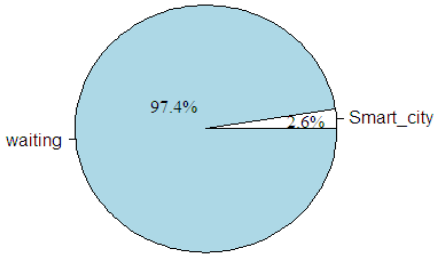
```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  194
```

cityy_test_labels	cityy_test_pred: Smart_city	waiting	Row Total
Smart_city	4	22	26
	0.154	0.846	0.134
	0.800	0.116	
	0.021	0.113	
waiting	1	167	168
	0.006	0.994	0.866
	0.200	0.884	
	0.005	0.861	
Column Total	5	189	194
	0.026	0.974	

Pie chart of Actual Smart cities and waiting cities



Pie chart of Predictied Smart cities and waiting cities



A) C5.0 Algorithm

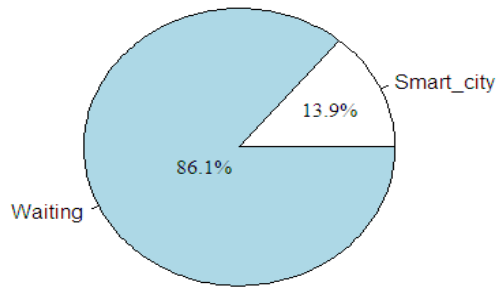
Cell Contents
N
N / Table Total

Total Observations in Table: 194

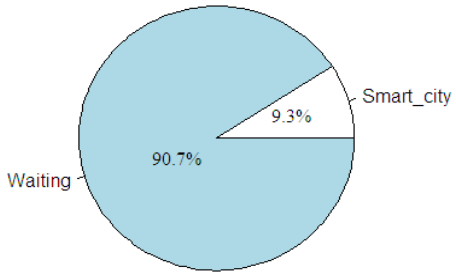
actual Smart cities and waiting cities	predicted smart cities and waiting cities: Smart_city	waiting	Row Total
Smart_city	10	17	27
	0.052	0.088	
waiting	8	159	167
	0.041	0.820	
Column Total	18	176	194

This Cross Table for C5.0 algorithm shows that a total of 176 of 194 predictions were true positive. Pie chats were created of both the actual smart cities and waiting cities, predicted smart cities and waiting cities.

Pie Chart of Actual Smart cities and waiting cities



Pie Chart of Predictied Smart cities and waiting cities



VII. CONCLUSION

In this study it was demonstrated that KNN algorithm is best suited for the analysis of population dataset. Experimental results show that error rate for KNN was low as compared to C5.0. In KNN algorithm, a total of 189 of 194 predictions were true positive that implies 97.4% of accuracy and in C5.0 algorithm, a total of 176 of 194 predictions were true positive that implies 90.7% of accuracy. This accuracy rate was achieved when the dataset had a total of 590 tuples out of which 194 tuples were treated as test data.

ACKNOWLEDGMENT

I thank Mr. Deep Kumar, Assistant Professor who guided me through my research-work.

REFERENCES

[1] Pantelis, Koutroumpis, and Leiponen Aija. "Understanding the value of (big) data." Big Data, 2013 IEEE International Conference on. IEEE, 2013.
[2] Al-Odan, Hussah A., and Ahmad A. Al-Daraiseh. "Open Source Data Mining tools." *Electrical and Information Technologies (ICEIT), 2015 International Conference on*. IEEE, 2015.
[3] Lohr, Steve. "The age of big data." New York Times 11.2012 (2012).
[4] Al Nuaimi, Eiman, et al. "Applications of big data to smart cities." Journal of Internet Services and Applications 6.1 (2015): 25.
[5] Caruana, Rich, and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
[6] Nugrahaeni, Ratna Astuti, and Kusprasapta Mutijarsa. "Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification." *Technology of Information and Communication (ISemantic), International Seminar on Application for*. IEEE, 2016.
[7] Fathima, Shameem A., and Nisar Hundewale. "Comparitive Analysis of Machine learning Techniques for classification of Arbovirus." *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*. IEEE, 2012.
[8] Venkatadri, M., and Lokanatha C. Reddy. "A review on data mining from past to the future." International Journal of Computer Applications 15.7 (2011): 19-22.
[9] Cios, Krzysztof J., and G. William Moore. "Uniqueness of medical data mining." Artificial intelligence in medicine 26.1 (2002): 1-24.
[10] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.
[11] http://censusindia.gov.in/Dist_File/datasheet-2923.pdf
[12] https://www.kaggle.com/bazuka/census2001
[13] D r. K. Venugopala Rao "Geo-informatics for Smart Cities- Indian Perspective" NRSC ISRO, Department of Space, Govt. of India Hyderabad, 2016