SVM and Cross-Validation using R Studio

Nainika Kaushik, Manjot Kaur Bhatia, Sonali Rastogi

Abstract: Each passing day data is getting multiplied. It is difficult to extract useful information from such big data. Data Mining is used to extract useful information. Data mining is used in majorly all fields like healthcare, marketing, social media platforms and so on. In this paper, data is loaded and preprocessed by dealing with some missing values. The dataset used is of Airbnb, the platform used for lodging and tourism industry. Analyzing the data by plotting correlation using spearman method. Further, applying PCA and Support Vector Machine classification technique on the dataset. There are various applications of SVM, it is used in face-detection, text and hypertext categorization, classification of images, bioinformatics and so on. SVM has high dimensional input space, sparse document vectors and regularization parameters therefore it is appropriate to use SVM. Cross-validation gives more accurate result. The dataset is divided into folds. The end product is the test set which is similar to full dataset. Confusion matrix is evaluated, grid approach is followed for building the matrix at various seeds and kernels (RBF, Polynomial). The aim of this research is to see which is the best kernel for the dataset.

Keywords: Big Data, Data mining, Machine learning Rattle, RStudio, Support Vector Machine.

I. INTRODUCTION

SVM is a supervised machine learning model used for classification. SVM is used in text and hypertext categorization. It is also used for classification of newsfeed as – sports, entertainment and politics. [1] SVM has various extensions like support vector clustering (SVC), Multiclass SVM [2], Transductive SVM, Structured SVM, etc. some common use of support vector machine is face detection text and hypertext categorization of classification of images bioinformatics protein food and remote homology detection, handwriting recognition generalized predictive control GPC. Airbnb is one of the most popular online market places where people book accommodation. Airbnb provides a platform to the people, offering lodging, hospitality and homestays. Support Vector Machine technique classifies rooms, on the basis of the type and rooms can be classified as entire home, private room, shared room. Support vector machine and K-means clustering is applied on the dataset and then to analyze it using a grid approach. Results are calculated at various kernels with different cost, seed value, degree parameters and correlation are plotted using spearman method.

II. RELATED WORK

Many authors have proposed work on predictive analysis using different techniques.

Jarou (2009) applied support vector machine and skin colour models for adult image detection[3]. Support vector machine is used in areas of facial detection but it is a cumbersome model. Skincare models use skin ratio, which is calculated using RGB information. A Hybrid scheme that is the combination of support vector machine and skin colour models is used to detect adult images. M. Dave (2009) proposed a parameter which acted as a churn, in the telecom industry. The decision tree model was built using R Studio. And from the computations, “gender” parameter was excluded, as it did not have any major impact on the data[4]. Wind power has become the most crucial parameter in this global warming environment. Wind power is free of cost and renewable source. Mathew S. (2019) worked on the dataset of NREL (National Renewable Energy Laboratory) using R Studio, and analysed various models on parameters- wind speed, wind direction, air temperature, air pressure and air density[5]. It was able to forecast one hour ahead. Nowadays, WhatsApp is the most usable application on our mobile phones[11]. It’s the first and the last thing we check during our day[6]. WhatsApp messages can be used analyze sentiments of people and to control circulation of fake news. Joshi S. (2016), performed sentiment analysis on WhatsApp chat database using R Studio. Wang Y. (2020) analyzed twitter data with 5,208 tweets containing hashtags such as #distracteddriving, #textanddrive, #textinganddrive and so on. SVM has high dimensional input space, sparse document vectors and regularization parameters therefore it is appropriate to use SVM. Cross-validation gives more accurate result. The dataset is divided into folds. The end product is the test set which is similar to full dataset. Confusion matrix is evaluated, grid approach is followed for building the matrix at various seeds and kernels (RBF, Polynomial). The aim of this research is to see which is the best kernel for the dataset.

III. PROPOSED METHODOLOGY

Dataset

Loading

Pre-Processing

Analyzing the dataset

(Plot correlation using spearman method)

Applying PCA and K-means clustering

Support Vector Machine

Confusion Matrix

Grid

Output

Figure 1: System Workflow

Revised Manuscript Received on September 05, 2020.  
Nainika Kaushik, Assistant Professor, Department of Information Technology, Jagan Institute of Management Studies - JIMS Rohini, Delhi, India.  
Dr. Manjot Bhatia, Professor, Jagan Institute of Management Studies - JIMS Rohini, Delhi, India.  
Sonali Rastogi, Computer Science Engineering Graduate, Jagannath University Bahadurgarh, Haryana, India.
In the proposed work dataset is loaded into the R Studio. Then pre-processing takes place in which data is cleaned. All the missing values need to be treated. The next phase is analyzing the dataset by plotting correlation using spearman method. Further applying PCA technique and K-means clustering algorithms. The last step is to implement SVM and compute confusion and grid approach. Figure 1 depicts the steps of the proposed work.

**Data Collection and Description**

The database of the Airbnb was fetched for 500 records for the analyzation process, it consists of the following parameters or attributes such as the house name, hostname, location, room type, number of reviews given, etc.

**Data Pre-processing**

Clean or pre-process the data to deal with missing values, mean was computed and entered into missing value where datatype was numerical. Otherwise for categorical data, deleted the obsolete row. So, a cleaned dataset of 500 records was created.

**Support Vector Machine**

SVM is used in both solving classification and regression problem statements. In linear separable system, a hyperplane is created to classify datasets into groups. Two parallel lines are created parallel to hyperplanes known as marginal planes. These marginal planes pass through nearest points of the groups. The distance between the two marginal planes is referred to as marginal distance. The points passing through marginal planes are called support vectors. In non-separable systems, SVM kernels are used. SVM kernels convert low dimensional systems to high dimensional systems. Always create a generalized model for better accuracy and the aim is to maximize marginal distance.

**Software Tool Used R**

R is an IDE, that is used for statistical computing and graphics. It can work with many operating systems such as Windows NT, macOS, Ubuntu, Fedora, Red hat Linux. The languages that are used are Java, C++ and JavaScript. Rattle is a Data mining and Data modelling statistical computing environment. One can download it from the comprehensive R archive network (CRAN) repository. R studio has only one data mining interface that is Rattle.

**R: Rattle package**

Include RGtk2 based graphical user interface. It allows to load CSV files. It provides functionality for data mining. To install Rattle by writing down two commands in R prompt.

`>library(rattle)`

`>rattle()`

Build number of models such as k means, tree, forest, support vector machine, Neural networks, etc.

---

**IV. PROPOSED METHODOLOGY**

The dataset was uploaded in the R. Studio. Rattle is used for GUI analyses. The .csv file was loaded, and partition was done as 70% training and 30% testing.

**Attributes in Dataset**

The attributes which are there in the data set are mentioned here.

| S.No. | Attribute Name |
|-------|----------------|
| 1     | Id             |
| 2     | Name           |
| 3     | host_id        |
| 4     | host_name      |
| 5     | neighborhood_group |
| 6     | Neighborhood   |
| 7     | Latitude       |
| 8     | Longitude      |
| 9     | room_type      |
| 10    | Price          |
| 11    | minimum_nights |
| 12    | Number_of_reviews |
| 13    | last_review    |
| 14    | reviews_per_month |
| 15    | calculated_host_listings_count |
| 16    | availability_365 |

**Loading the dataset**

Select “Data” tab on that load the .csv file, then change the partition to 70% training and 30% testing. Click on “Execute” all the attributes will be shown with its data types, input, target, risk, ident, ignore, weight, comments parameters. Select target variable, selected room_type as the target variable as it is categoric. SVM can be applied only on categoric attribute.

![Fig. 2. R Studio Interface](image-url)

**Fig. 3. Attributes in the Dataset with Target Variable**

**Summary of the dataset**

In “Explore” tab, selecting “Summary”, All the details of the dataset will be visible in detail such as datatype and statistical information such as Min, First Quartile, Median, Mean, Third Quartile and Max., for each attribute[8].
Describing the data

Under the explore tab and summary radio button by selecting the describe option, it shows all the attributes with the available observations and missing observations of the attributes along with their statistical interpretation such as Mean, Median etc.[9].
Fig. 6(b). describes the statistical data about the attributes as: Minimum Nights, Number of reviews and Reviews per Month.

Similarly, under the explore tab and summary button by selecting the skewness option, it shows the skewness of all the variables. The skewness tells about the tail of the distribution of the data. The skewness measure of variables can be positive, negative and as well as undefined.

Fig. 6(b). Basics of Attribute

Fig. 6(c). describes the statistical data about the attributes as: Reviews per Month, Calculated Host Listing Counts.

Fig. 6(c). Basics of Attribute

Fig. 6(d). describes the statistical data about the attributes as: Calculated Host Listing Counts, Availability_365 (i.e. Throughout the year).

Fig. 6(d). Basics of Attribute

Knowing more about the Data
Under the explore tab and summary button by selecting the kurtosis option, it shows the peakness of all the variables around the mean. The kurtosis measure of variables can be positive and as well as negative.

Fig. 7. Kurtosis of Attributes

Similarly, under the explore tab and summary button by selecting the skewness option, it shows the skewness of all the variables. The skewness tells about the tail of the distribution of the data. The skewness measure of variables can be positive, negative and as well as undefined.

Fig. 8. Kurtosis of Attributes

Under the explore tab and summary radio button by selecting the show missing checkbox, missing value summary shows up.

Fig. 9. Missing Summary of Attributes

Under the explore tab and correlation button by selecting one of the methods among listed as spearman, Karl Pearson and many more. The correlation tells about the statistical dependence of the random variables.
Selecting the appropriate method, the dependence among the variables is shown the below Graph.

Fig. 10(a). Ordered Correlation Graph through Spearman’s Rank Method
Fig. 10(b). Ordered Correlation Graph through Pearson’s Method

Fig. 10(c). Ordered Correlation Graph through Kendall’s Method

Fig. 10(d). Ordered Correlation Clusters through Pearson’s Method

Fig. 10(e). Ordered Correlation Clusters through Kendall’s Method

Fig. 10(f). Ordered Correlation Clusters through Spearman’s Method

Principal Components About Data

Principal Components Analysis (PCA) is a method to convert the correlated variables or said ‘attributes’ into a smaller number of uncorrelated variables or attributes. PCA is only applicable to numerical components and so cannot be applied to categorical data.

Fig. 11(a) describes the PCA through SVD method for the dataset.

Fig. 11(b). PCA by SVD Method

Fig. 11(c). describes the plotted graph among the principal components of dataset.
SVM and Cross-Validation using R Studio

Fig. 11(c). Principal Components of Attributes

Fig. 11(d). describes the PCA through Eigen method for the dataset.

Fig. 11(d). PCA by Eigen Method

Fig. 11(e). PCA by Eigen Method

Fig. 11(f). describes the plotted graph among the principal components of dataset.

Fig. 11(f). Principal Components of Attributes

Fig. 11(g). describes the plotted graph for the principal component’s importance of dataset with variance through Eigen Method.

Fig. 11(g). Principal Components of Attributes

Fig. 11(h). describes the plotted graph for the principal component’s importance of dataset with variance through SVD Method.

Fig. 11(h). Principal Components of Attributes Testing the Data

Test the data using the prebuild testing methods in rattle. Each testing technique has different basis for judging the technique. Such as they test the data on basis of distribution of data, location of the average, variation in the data, correlation.

Fig. 12. Testing the Attributes Transform the data

Rescale the data, it shows the details of all the variables, its datatype and number missing. Clean the data at this stage, remove all the missing values.

Fig. 13. Rescaling of Attributes
Cluster Formation
Under the Cluster tab by selecting the suitable model for clustering such as Kmeans in the below picture. Necessary details such as number of clusters to be formed, seeds, runs, re-scaling option, Use HClust Centers, Iterate Clusters, stats tab, and plotting tab: data, discriminant, weights have to be entered accordingly to the dataset entered.
Kmeans method is used for cluster formation and enter details such as required clusters as three, setting the seed level and number of runs and rescaling it.

![Fig. 14(a). Clustering by Kmeans](image1)

![Fig. 14(b). Clustering by Kmeans](image2)

Association of Data
Association Analysis is widely used in Data Mining. Rattle is supporting Association Mining rules through the tab of “Association” of unsupervised paradigm. It is based on market basket analysis and has two dependent factors as Support and Confidence. Either ident or Target Variable is used by rattle for the association analysis. Retail business widely use this approach for their purpose.

![Fig. 14(c). Association Rule Analysis](image3)

Model Selection
Here under the model bar select the suitable type of model and set the required parameters such as kernel, degree etc.

![Fig. 15(a). SVM by Radial Basis (rbfdot)](image4)

![Fig. 15(b). SVM by Polynomial (POLYDOT) with degree =1](image5)

![Fig. 15(c). SVM by Polynomial (POLYDOT) with degree =2](image6)
Evaluating the data
Here under the Evaluate bar choose the type and model and select the data mode i.e. testing, training, etc., risk variable and report type. Note that the error matrix generated will be different for each seed and depending upon the model chosen.

The evaluation tab basically is used for the evaluation of the performance of the models chosen for the selected dataset. Evaluation tab offers various types of evaluation models such as Error matrix, risk, cost curve, hand, left, ROC, sensitivity, Pr v Ob, Score and models that are available in the evaluation tab follows as- decision tree, Ada boost algorithm, random forest algorithm, SVM, Neural networks, survival, K means, HClust[10].

To select the type of evaluation for the selected data set one can choose among the types of radio buttons shown. The model for the evaluation can be chosen by selecting the appropriate checkbox. Also, for the data one has to select the radio button among the listed ones. Also select the appropriate report button among the listed ones.

Fig. 16(a) and Fig.16(b) display the error matrix for the dataset. The error matrix displays the predicted as well as the overall error along with the average class error.

Grid Approach
Grid matrix is a structured approach of displaying the data in a two-dimensional format. One can change the size of a matrix accordingly to the data set or dynamically. One can form grid Matrix in R studio by importing a package of grid layout function[12].

Seed is a vector in R that stores random number that are declared by normal generator for random number generator RNG. Useful in reproducing random objects and creating simulations.

In the Table 2, it can be clearly seen that 5 splits which corresponds to different seeds value. Here data set is analyzed by using various different kernels, degree and cost parameters. It can be seen that polynomial kernel gives the least overall average error (18.64%) with degree as 2, cost parameters as 1 and 10.

| Split | Seed | RBF C=1 | RBF C=10 | Poly D=1, C=1 | Poly D=1, C=10 | Poly D=2, C=1 | Poly D=2, C=10 |
|-------|------|---------|----------|---------------|----------------|---------------|----------------|
| Split 1 | 42   | 20.70%  | 22.10%   | 23.50%        | 23.50%         | 21.40%        | 21.40%         |
| Split 2 | 255107 | 24.10%  | 21.50%   | 18.70%        | 18.70%         | 22.10%        | 22.10%         |
| Split 3 | 894393 | 13%     | 13.60%   | 15%           | 15%            | 11.50%        | 11.50%         |
| Split 4 | 114762 | 19.30%  | 21.40%   | 22.10%        | 22.10%         | 21.40%        | 21.40%         |
| Split 5 | 307054 | 20.20%  | 16.10%   | 17.50%        | 17.50%         | 16.80%        | 16.80%         |
| Average |      | 19.46%  | 18.94%   | 19.36%        | 19.36%         | 18.64%        | 18.64%         |

V. CONCLUSION
Paper analyses a dataset of 500 records in this support vector machine model and K-means clustering is applied. The dataset is first cleaned and then confusion matrix is evaluated. Then grid approach is applied at different seeds. Various kernels are evaluated such as polynomial, radial basis at different degree and cost parameters for SVM. Observed that the best kernel for the dataset is polynomial with degree 2 and cost parameters as 1,2.

REFERENCES
1. Mayor, Shweta & Pant, Bhasker. (2012). Document Classification Using Support Vector Machine. International Journal of Engineering Science and Technology.
2. Duan, Kai-Bo; Keerthi, S. Sathiya (2005). "Which Is the Best Multiclass SVM Method? An Empirical Study" (PDF). Multiple Classifier Systems. LNCS. 3541.pp. 278-285. CiteSeerX 10.1.1.110.6789. doi:10.1007/11494683_28. ISBN 978-3-540-26306-7.
3. B. Choi, B. Chung and J. Ryoo, "Adult Image Detection Using Bayesian Decision Rule Weighted by SVM Probability," 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, 2009, pp. 659-662.

4. Vani Kapoor Nijhawan, Mamta Madan, Meenu Dave (2009). A Comparative Analysis Using RStudio for Churn Prediction. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7S2, May 2019

5. Pawar A., Jape V.S., Mathew S. (2019) Wind Power Forecasting Using Support Vector Machine Model in RStudio. In: Mallick P., Balas V., Bhai A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore

6. Joshi S. (2019) Sentiment Analysis on WhatsApp Group Chat Using R. In: Shukla R., Agrawal J., Sharma S., Singh Tomer G. (eds) Data, Engineering and Applications. Springer, Singapore

7. Qian C., Li Y., Zuo W., Wang Y. (2020) Analysis of Driving Safety and Cellphone Use Based on Social Media. In: Stanton N. (eds) Advances in Human Factors of Transportation. AHFE 2019, Advances in Intelligent Systems and Computing, vol 964. Springer, Cham

8. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2017). URL https://www.R-project.org/

9. Zhao, Y.: R Reference Card for Data Mining. http://www.Rdatamining.com Online. Access 10 June 2017

10. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Cambridge (2016)

11. Patil, S.: WhatsApp group data analysis with R. Int. J. Comput. Appl. 154(4) (2016)

12. Torgo, L.: Data Mining with R Learning with Case Studies. CRC Press, Taylor & Francis Group an Informa Business (2011)

AUTHORS PROFILE

Nainika Kaushik She is working as an assistant professor in department of information technology, JIMS. She has completed her MTech from IGDTUW with the specialization in mobile and pervasive computing and B. tech from GGSIPU in computer science and engineering. Her areas of interest are data mining, mobile computing and software testing. She has published many papers in international journals.

Manjot Kaur Bhatia She is working as a professor in department of information technology, JIMS. She has completed her M.C.A., M.Phil. and Ph.D. from University of Delhi. Her research topic is in the area of Information Security. She has more than 20 years of teaching experience. She is actively involved in teaching and research in the areas of Security, Databases, Linux, Operating System and Information Security. Her other areas of interest include Cloud Computing, Steganography, Data Hiding, and Software testing. She has been guiding many Ph. D. scholars in their research work and motivated MCA students also towards the research work. She has published many research papers in various refereed International journals and has presented number of papers in international conferences. She has been the session chair in various international conferences.

Sonali Rastogi She has completed her B. Tech from Jagan Nath University, Bahadurgarh (JIMS) with the specialization in computer science and engineering. She has recently published a paper in international journal. Her areas of interest are influential marketing, data mining, and software testing.