Utilizing Genetic Algorithm for Optimal Feature Selection in Multiclass Classification Problems: A Comprehensive Study

Gampala Sivani

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *The Crucial Role of Component Determination in Information Preprocessing for Data Mining: An Exploration into Genetic Algorithm Based Feature Selection for Multiclassification*

The process of component determination stands as a pivotal initial step in information mining endeavors. At its core, feature selection aims to identify a subset of potential features by excluding those with minimal predictive value as well as highly correlated redundant features. This paper proposes the application of a genetic algorithm tailored specifically for multiclassification tasks in the context of feature selection. Initially, the similarities among features are computed as part of the feature selection process. Subsequently, the reduced feature set is employed for multiclassification tasks.

Multiclass representation presents a significant challenge within the realm of model validation, involving the assignment of input models into one of several classes. Given the prevalence of class overlap in real-world scenarios, the multiclass assignment task becomes notably more complex and demanding compared to binary classification. To investigate this, our study focuses on the solar dataset sourced from the University of California at Irvine Machine Learning Data Repository.

Our findings highlight the efficacy of Support Vector Machine (SVM) and Multilayer Perceptron (MLP) in achieving notable performance during the classification phase. The multiclass representation approaches demonstrate promising results when compared to solutions obtained using conventional classifiers and those leveraging shared information.

I. INTRODUCTION

Multi-name plan is an AI request task that contains various classes, or yields. AI gathering is the way toward approximating the arranging limit that maps the data test to target class/name [1]. In standard portrayal issues, the data tests identify with only one target mark. This sort of game plan is called single-mark request [2]. Twofold request incorporates describing the data tests into both of two sets subject to a specific portrayal metric. The number of disjoint names is 2 for twofold plan. There are a couple of genuine application issues including distinctive target marks achieving the improvement of multi-class game plan [8]. Multi-class gathering incorporates organizing the data tests into various classes. Character affirmation, biometric recognizing evidence and security, face affirmation is a part of the application spaces of multi-class course of action [9].

In any case, in various authentic applications, the data tests contrast with various target names. This condition of portrayal, where the data identifies with a lot of class stamps as opposed to one, is called multi-name gathering. Multilabel plan has become a rapidly emerging field of AI in view of the wide extent of utilization spaces and the omnipresence of multi-name issues in veritable circumstances [8].

So, to perform gathering tasks, all judicious request models don't maintain multi-class portrayal like Logistic backslide, support Vector Machine as those are planned to perform Binary course of action and don't maintain request tasks numerous classes. Strangely, Decision tree gathering, K-nearest neighbor, Naive Bayes Classification and neural association-based models give predominant execution for Multi-Class Classification.

II. FEATURE SELECTION

Highlight determination has been broadly explored and utilized by the AI and information mining local area. In this unique situation, a component, likewise called quality or variable, addresses a property of a cycle or framework than has been estimated or built from the first information factors. The objective of highlight choice is to choose the littlest feature subset given a specific speculation blunder, or alternatively finding the best element subset with k features, that yields the base speculation mistake [3]. Extra destinations of feature determination are as per the following: (I) further develop the speculation execution as for the model assembled utilizing the entire arrangement of highlights, (ii) give a more vigorous speculation and a quicker reaction with inconspicuous information, and (iii) achieve a superior and less complex comprehension of the interaction that creates the information [7]. Highlight determination methods are normally characterized in three principle gatherings: covering,

inserted, and channel techniques. Coverings utilize the acceptance learning calculation as a component of the capacity assessing highlight subsets. The presentation is normally measured in terms of the grouping rate got on a testing set, i.e., the classifier is utilized as a black box for evaluating highlight subsets [4]. Albeit these strategies may accomplish a decent speculation, the computational expense of preparing the classifier a combinatorial number of times becomes prohibitive for high-dimensional datasets.

2.1 Genetic Algorithm (GA)

Genetic Algorithms for Feature Selection Among the different classifications of highlight determination calculations, the transformative calculations especially GAs are mainstream and generally utilized. Genetic calculations are search calculations dependent on the standards of regular determination and hereditary qualities, presented by J Holland in the 1970's and propelled by the natural development of living creatures. Genetic calculations theoretical the issue space as a populace of people, and attempt to investigate the fittest individual by delivering ages iteratively. GA develops a populace of introductory people to a populace of great people, where every individual addresses an answer of the issue to be settled. The nature of each standard is estimated by a wellness work as the quantitative portrayal of each standard's transformation to a specific climate. The methodology begins from an underlying populace of haphazardly produced people. During every age, three essential hereditary administrators are consecutively applied to every person with specific probabilities, for example determination, hybrid and change [6].

In GA, the inquiry space comprises of strings, every one of which addressing a competitor answer for the issue and are named as chromosomes. The target work worth of every chromosome is called its wellness esteem. Populace is a bunch of chromosomes alongside their related wellness. Ages are populaces produced in an emphasis of the GA [10]. Hereditary calculation to look through a space of up-and-comer answers for recognize the best one is as per the following:

2.2 GA Steps

1. Generate arbitrary populace of n chromosomes (reasonable answers for the issue).
2. Evaluate the wellness f(x) of every chromosome x in the populace.
3. Create another populace by continuing after strides until the new populace is finished
    a. Select two parent chromosomes from a populace as indicated by their wellness (the better wellness, the greater opportunity to be chosen).
    b. With a hybrid likelihood gets over the guardians to frame another posterity (youngsters). On the off chance that no hybrid was performed, posterity is a precise of guardians.
    c. With a change likelihood transforms new posterity at every locus (position in chromosome).
    d. Place new posterity in another populace.
4. Use new produced populace for a further run of calculation.
5. If the end condition is fulfilled, stop, and return the best arrangement in current populace.
6. Go to stage 2.

III. METHODOLOGY

This segment gives the succinct considered picked regulated models of Support Vector Machine and Multilayer Perceptron.

3.1 Support Vector Machine (SVM)

SVM is a learning calculation that is being utilized for relapse and grouping learning assignments. The dataset focuses are addressed in space in SVM and are isolated into focuses and bunches with comparable constructions that fall into similar gatherings [13]. The information is viewed as p-dimensional for direct SVM that can be parceled by the size of p-1 planes known as hyper planes [14]. Accordingly, the planes partition the arrangement of limits and information space among the information bunches for relapse or order learning task. The most ideal choice is chosen dependent on the distance between the

partitioned classes. In this manner, the plane with the most elevated cutoff between these two classifications is called greatest edge hyperplane.

3.2 Artificial Neural Network (ANN)

ANN mirrors the capacities and exercises of the mind of person which is recognized as the hubs, which is actually known as or called counterfeit neurons [11]. The neurons impart and communicate information and data among themselves in type of 0 s and 1 s or mix and every neuron has a particular weight given to it, which shows its capacities and jobs to carry out in the framework [4]. The construction of ANN is partitioned into layers, directly from information gathering layer, input layer, center or secret layer to yield layer which is called extraction or grouping layer. Each layer has a particular capacity to perform and change information into the significant data to get a definitive and ideal outcome [5]. The Activation and move work assume a basic part in the exercises do by neurons.

3.2.1 Multilayer Perceptron (MLP)

MLP stands for Multilayer Perceptron, which is a type of artificial neural network (ANN). It consists of multiple layers of nodes (or neurons), including an input layer, one or more hidden layers, and an output layer. Each node in one layer is connected to every node in the subsequent layer, and each connection is associated with a weight.

MLPs are widely used for various machine learning tasks, including classification, regression, and pattern recognition. They are known for their ability to learn complex relationships in data and can be trained using algorithms such as backpropagation.

A MLP is a hero among the most by and large saw Neural Network plan that has been utilized for different applications. The MLP sort out is ordinarily made out of various focuses or managing units, and it is sorted out into a development of no under two layers [5]. The fundamental layer (or the most lessened layer) is named as a data layer where it gets the outer data while the last layer (or the most astounding layer) is a yield layer where the reaction for the issue is gotten. The hidden layer is the broadly engaging layer in the information layer and the yield layer, and may outline with somewhere near one layers. The plan of MLP could be imparted as a nonlinear improvement issue. The target of MLP learning is to track down the best loads that limit the separation between the data and the yield. The most prevalent preparing assessment utilized in NN is Back engendering (BP), and it has been utilized in managing different issues in model assertion and depiction. This calculation relies several limits, for example, unique covered focus focuses at the concealed layers learning rate, energy rate, authorization work and the amount of planning to occur.

IV. EXPERIMENTAL RESULTS

This part depicts the test results got by applying the proposed multi-name characterization calculation to a Solar dataset are taken from the UCI AI storehouse [12]. In the Solar dataset, there are 323 records, 13 credits and 6 class marks are displayed in the figure-1. We have utilized the weka to analyze our proposed calculations. Weka is a state-of-the-art office for making ML techniques and their application to genuine data mining issues.

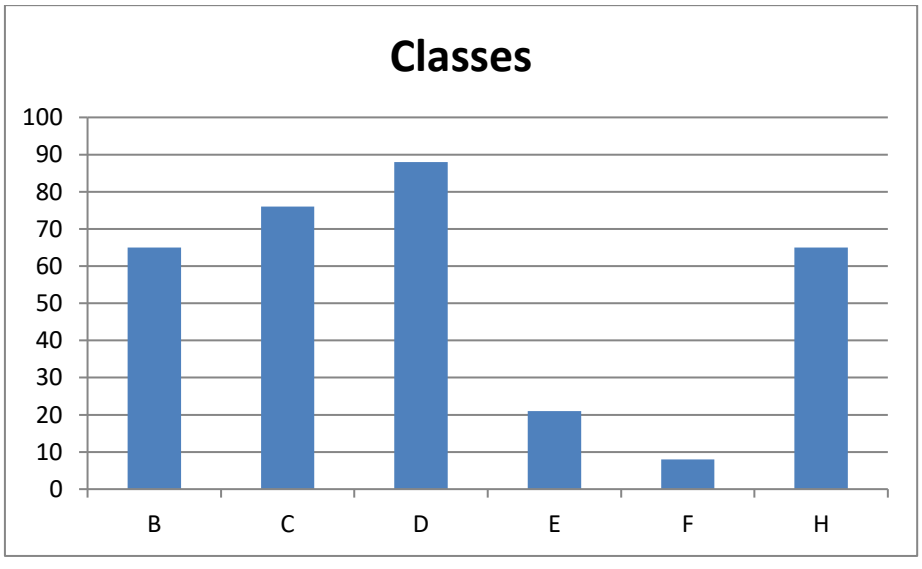

Figure-1: Class-wise distribution of labels of Solar dataset

The Solar dataset detailed information and summary of statistical analysis as shown in the figure-2 and figure-3.

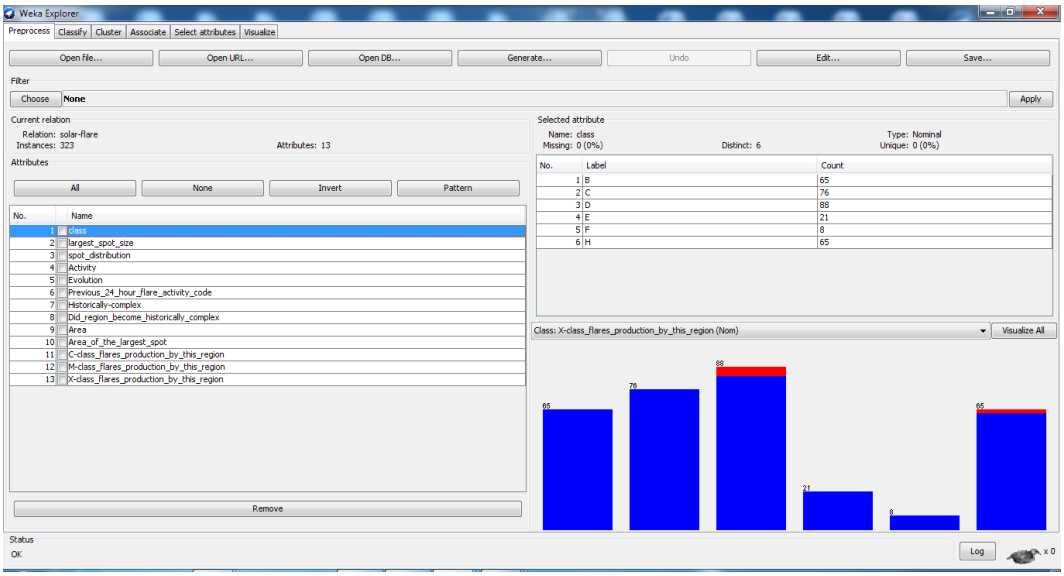

Figure-2: Solar dataset details

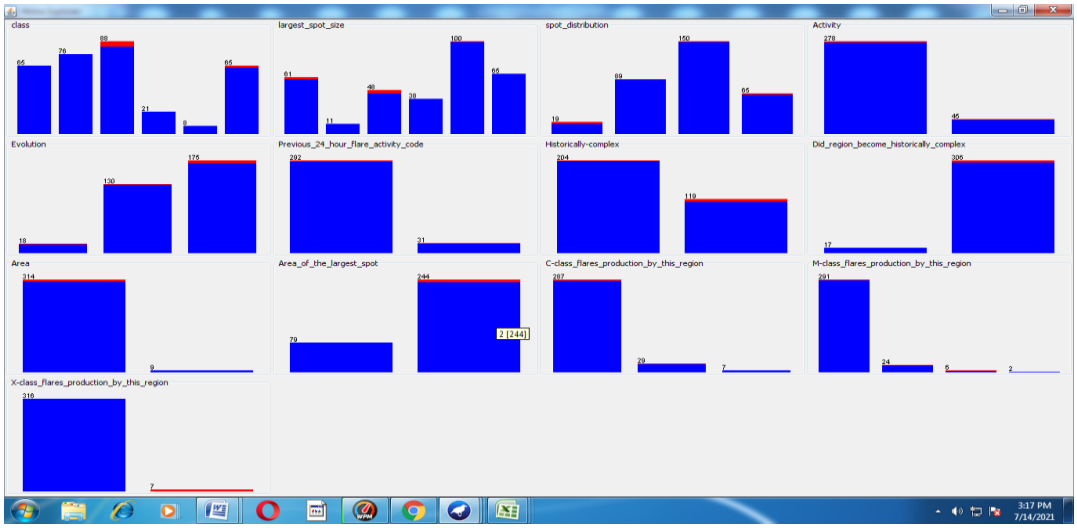

Figure-3: Statistical Summary of Solar Dataset

4.1 Results

In the first stage MLP and SVM algorithms are trained on the original set of features was used in the experiment. In the second stage we implement a GA algorithm for obtaining the adequate number of features to identify the features selected. The results that we got for MLP and SVM without GA based feature selection and with GA based feature selection are shown below in the table-1 and same as shown in the figure-4 with their corresponding values.

TABLE-1
PERFORMANCE OF CLASSIFIERS

Algorithm	Accuracy	precision	Recall
SVM without GA	95	95	95
SVM with GA	98	98	98
MLP without GA	95	94	94
MLP with GA	97	97	97

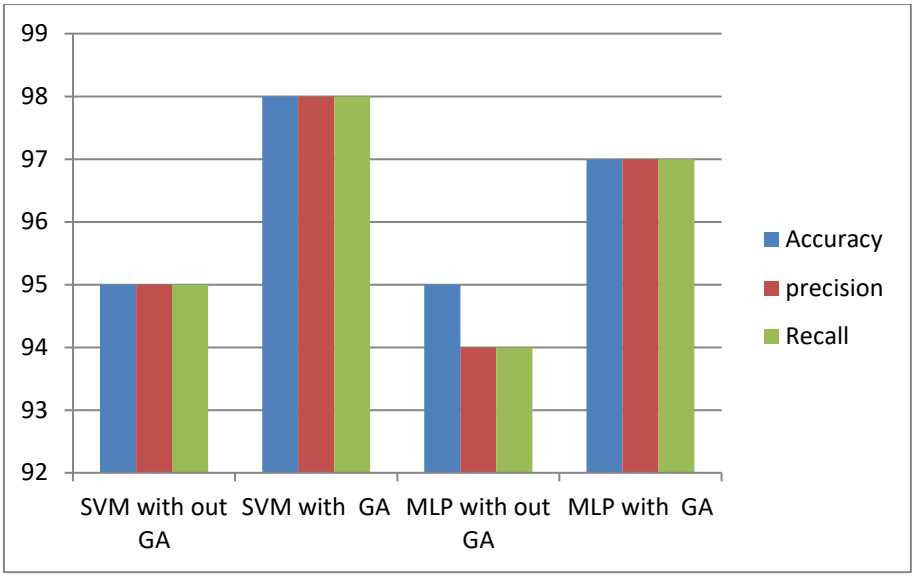

Figure-4: Performance of Classifier Multi- Label Classification

From the figure-4, we observe the performance of SVM without GA based on accuracy has got 95%, whereas the performance of SVM with GA feature selection based on accuracy has achieved 98%. However, there is an improvement in the accuracy with feature selection. The accuracy rate is increased 3% with feature selection.

Also, we observe the performance of MLP without GA based on accuracy has got 95%, whereas the performance of MLP with GA feature selection based on accuracy has achieved 97%. However, there is an improvement in the accuracy with feature selection. The accuracy rate is increased 2% with feature selection.

In our experimental result the SVM with GA feature selection algorithm shows the highest accuracy compared with MLP with GA. With the improvement the accuracy, the proposed model demonstrated that it performs well after selecting relevant features. This result provided new insight using a classification learning algorithm and reduction technique to selection relevant and important feature in order to improve the accuracy of the system and to identify possible features which may contribute to this improvement. Most of the proposed research system could effectively utilize feature selection process to improve detection rate of their system and minimize considerably the false alarm rate.

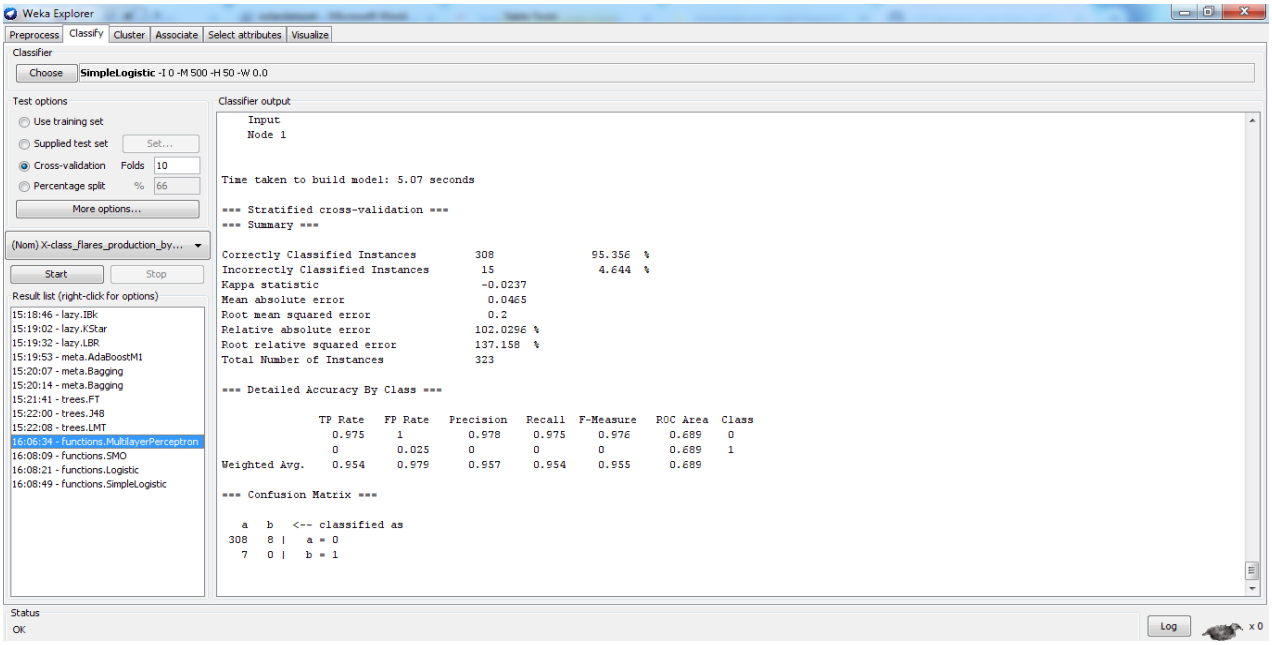

Figure-5:Screen Shot of Results

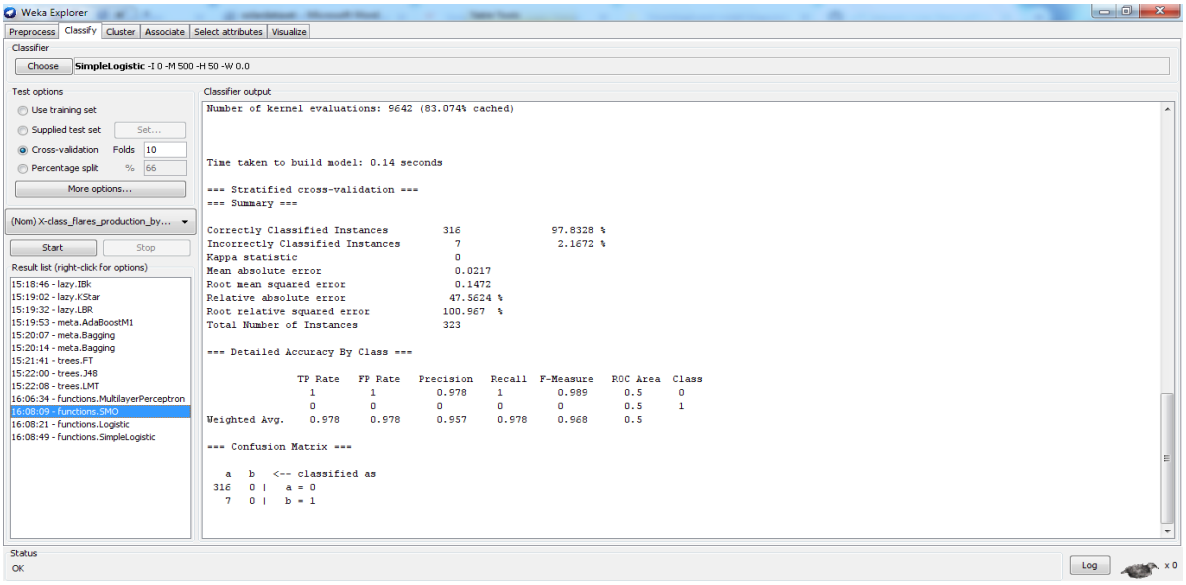

Figure-6: Screen Shot of Results

V. CONCLUSION

This paper explores effective solutions for feature selection in classification tasks. It introduces a feature selection method combining Genetic Algorithm (GA) for feature elimination with Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers. The integrated approach demonstrates promising classification results through effective feature selection, preprocessing, and classification techniques. Evaluation using various classification metrics confirms the method's effectiveness, showing improved model accuracy with reduced feature count. Additionally, the study emphasizes the importance of efficient detection algorithms and feature selection methods for large datasets

REFERENCES

[1] G. Bo and H. Xianwu, "SVM multi-class classification," Journal of Data Acquisition & Processing, vol. 21, pp. 334-339, 2006.
[2] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in Data mining and knowledge discovery handbook, ed: Springer, 2010, pp. 667-685.
[3] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
[4] H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2000)
[5] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann
[6] J. Holland, “Adaptation in natural and artificial systems”, University of Michigan press, Ann Arbor, 1975.
[7] Kohavi R, John GH (1997) Wrappers for feature subset selection. ArtifIntell 97(1–2):273–324
[8] M. Boutell, X. Shen, J. Luo, and C. Brown, "Multi-label semantic scene classification," technical report, dept. comp. sci. u. rochester2003.
[9] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, pp. 1819-1837, 2014.
[10] Noraini Mohd Razali, John Geraghty “A genetic algorithm performance with different selection strategies”, Proceedings of the World Congress on Engineering Vol II, 2011.
[11] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”, 2nd edition, Addison Wesley, 2005.
[12] UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets.html)
[13] Vapnik V.N, “Statistical learning Theory”, John Wiley and Sons, New York, USA, 1998.
[14] Vapnik V.N,”The Natural of Statistical Learning Theory, Springer-Verlag, New York, USA, 1995.