Journal of Positive School Psychology
2022, Vol. 6, No. 3, 2812–2816

http://journalppw.com

SVM Hyperplane Misclassification Control by Finding Optimum Cost of Misclassification with Boundary Value Analysis Technique

[1]Kumar Gaurav Kapoor, [2]Avadhesh Kumar

[1,2] Galgotias University, Greater Noida (U.P.),India
Email: [1]kumargaurav.kapoor@gmail.com, [2]avadheshkumar@galgotiasuniversity.edu.in

Abstract

In Health care domain roughly 80% of data in electronic medical records consists of physicians' unstructured notes. To unlock this important data we need a different approach than what we used to analyse structured data. That's one place where machine learning comes in and for this SVM (Support Vector Machine) is extensively used to identify the handwritten digits and words based on pixel given as features. SVM uses the concept of Hyper Planes which leads to a boundary which classifies the data set. Key area of the Research paper is to get Optimum Hyper plane.

Keywords— Support Vector Machine, Hyperplane, Slack Variable

I. INTRODUCTION

Support Vector Machine (SVM) is an advanced machine learning technique which has a unique way of solving complex problems such as image recognition, face detection, voice detection [1]. SVMs belong to the class of linear machine learning models. A line that is used to classify one class from another is called a hyperplane.

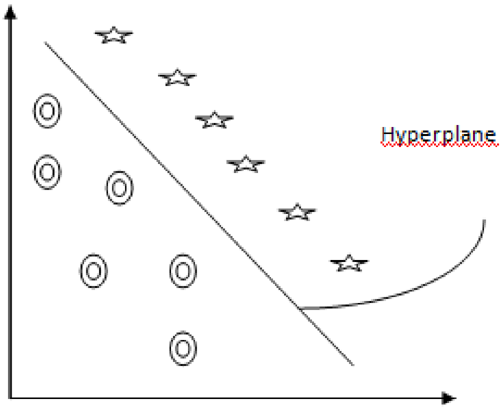


Figure 1: Depicting Hyperplane

A positive value (blue points in the plot above) would mean that the set of values of the features is in one class; however, a negative value (red points in the plot above)would imply it belongs to the other class. A value of zero would imply that the point lies on the line (hyperplane).

However, there could be multiple lines (or hyperplanes in general) possible, which perfectly separate the two classes. How to get the optimum hyperplane.is the key challenge.

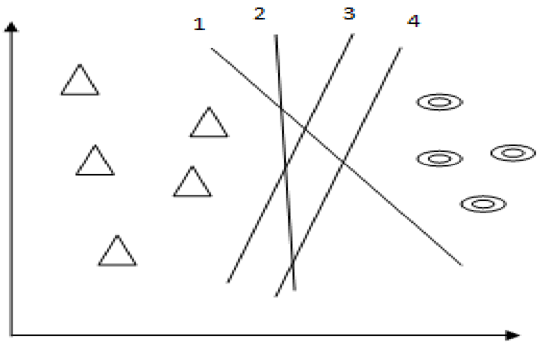


Figure 2: Choosing 1 or 2 or 3 or 4 is the biggest challenge in Hyperplane finding

Partially Intermingled data case create more confusion in finding hyper plane for example in below diagram.

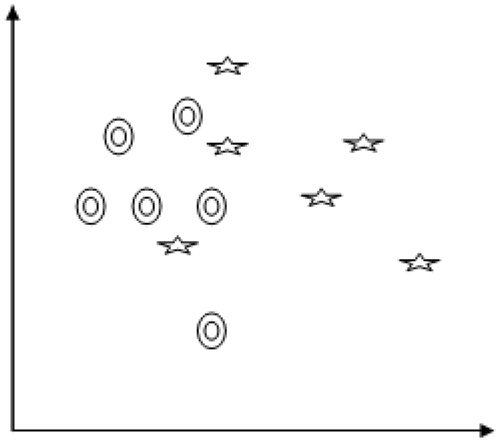


Figure 3:Partially Intermingled Data

It seems that the maximal margin line (hyperplane) is not even possible in this case. In this case, if want to create a linear hyperplane, we will inevitably need to misclassify a few data points. In other words, some points will need to fall on the wrong side of the hyperplane.

Slack variable: A slack variable is used to control misclassifications.It tellswhere an observation is located relative to the margin and hyperplane.

For points which are at safe distance from the hyperplane, the value of the slack variable is 0.

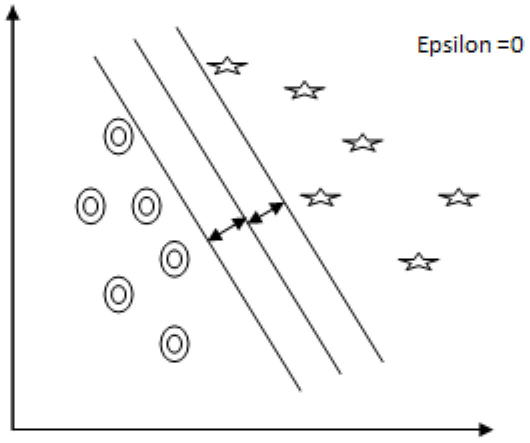


Figure 4:Slack Variable

On the other hand, if a data point is correctly classified but falls inside the margin (or violates the margin), then the value of its slack is between 0 and 1 [2, 3].

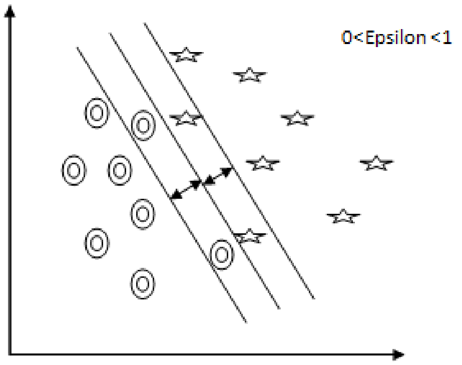


Figure 5 :Epsilon value lies between 0 and 1

Finally, if a data point is incorrectly classified (i.e. it violates the hyperplane), the value of epsilon > 1.

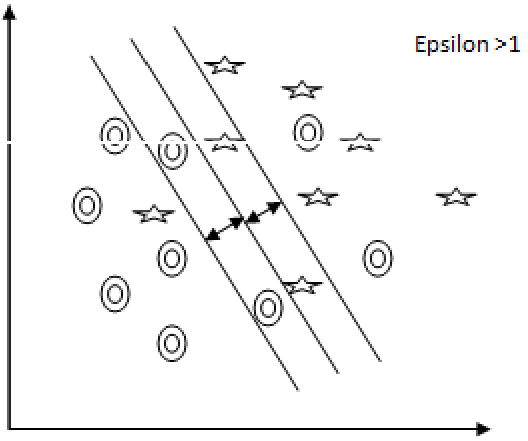


Figure 6 :Epsilon value> 1

Each data point has a slack value associated. The value of the slack lies between 0 and infinity . Lower values of slack are better than higher values.

The summation of all the epsilons of each data point is denoted by cost or 'C' [3]i.e.

$$\sum \epsilon i \leq C.$$

When C (summation of all the epsilons) is large, the slack variables can be large thus allowing a large number of data points to be misclassified or to violate the margin.

On the other hand, when C is small, it force the individual slack variables to be small, thus allowing many data points to fall on the wrong side of the margin or the hyperplane.

Finding optimum cost of misclassification is the main objective of the paper.

II. LITERATURE REVIEW

Health care industry is focusing on Digitization that leads to electronic health recording system. Today digitally means analysis of records which occurred due to use of sensors, remote monitoring, and apps to provide continuous data [4].

Now Health care industry started using Predictive analysis Predictive analysis Modelling uses machine learning algorithms, in which the machine learns from the data just like humans learn from their experiences. Machine learning can be used heavily in the industry. Let's go deeper into the types of models that come under machine learning[5].

Machine learning models can be classified into the followingthree types based on the task performed and the nature of the output:

1. Regression: The output variable to be predicted is a continuous variable, e.g. blood pressure of a patient
2. Classification: The output variable to be predicted is a categorical variable, e.g. classifying patients with fever or not.
3. Clustering: No pre-defined notion of label allocated to groups/clusters formed, e.g. patient segmentation

Machine learning models classified into two broad categories as follows:[7, 8].

1. **Supervised Learning Methods**
   - Past data with labels is used for building the model.
   - Regression and classification algorithms fall under this category.
   - The past data is divided into training and testing data sets for building the model.

2. **Unsupervised Learning Methods**[9]
   - No pre-defined labels are assigned to past data
   - Clustering algorithms fall under this category

Over fitting is a phenomenon where a model becomes too specific to the data it is trained on and fails to generalize to other unseen data points in the larger domain. A model that has become too specific to a training data set has actually 'learnt' not just the hidden patterns in the data but also the noise and the inconsistencies in the data. In a typical case of over fitting, the model performs very well on the training data but fails miserably on the test data.

A model should never be evaluated on data it has already seenb before. With that in mind there will either one of two cases –

1) The training data is abundant or
2) The training data is limited.

The first case is straightforward because it can use as many observations as like to both train and test the model. In the second case, however, it has to find some 'hack' so that the model can be evaluated on unseen data and at the same time doesn't eat up the data available for training. This hack, commonly used in statistics, is called cross-validation [4].

On dividing the available data into 75:25 or 70:30, themajority data is called the training set, the other part is called testing set [10].SVMs belong to the class of linear machine learning models which uses a linear function(i.e. of the form y = ax +b) to model the relationship between the input x and output y.

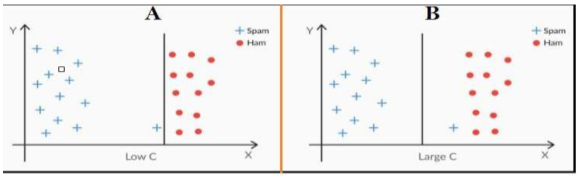
For example, in logistic regression, the log(odds) of an outcome (say, defaulting on a credit card) is linearly related to the attributes x1, x2, etc.

$$\log(\text{odds of default})=\beta0+\beta1X1+\beta2X2+...\beta nXn$$

Similarly, SVMs are also linear models [11].SVMs need attributes in the numeric form. To summarize, SVMs are basically a linear models that takes numeric attributes. But if the attributes are non-numeric, then have to convert them to a numeric form in the data preparation stage [1].

III.PROBLEM STATEMENT

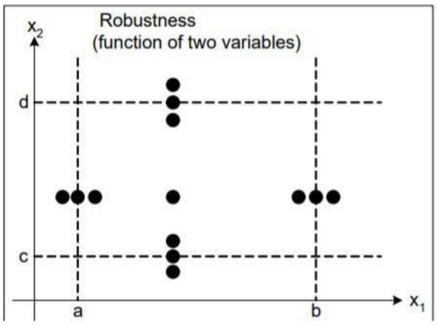
Support Vector Machine uses the concept of Hyperplanes which leads to a boundary which classifies the data set. Problem is how to get Optimum hyperplane with cost of misclassification so thus to perfectly separate classes [3].



If C is large, the slack variables (epsilons(ϵ)) can be large, i.e. you allow a larger number of data points to be misclassified or violate the margin; and if C is small, you force the individual slack variables to be small, i.e. you do not allow many data points to fall on the wrong side of the margin or the hyperplane [4, 6, 12].

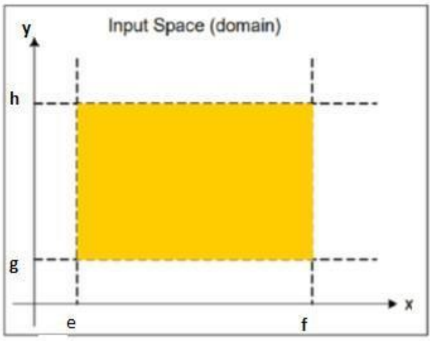
IV. PROPOSED SOLUTION

Application of BVA for finding Optimum Cost of Misclassification. Boundary Value Analysis (BVA) analysis the values which lies on the boundaries, values just above the boundary and just below the boundary [13].



Example: x and y
e <= x<= f
g <= y <=h

In this e and g are lower boundary f and h are upper boundary



• Min range----------	Minimum
• Min+ range --------	Just above Minimum
• Nom range --------	Nominal
• Max range ---------	Maximum
• Max- range --------	Just below Maximum

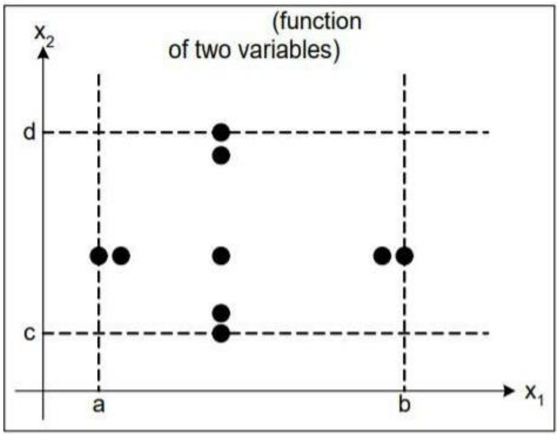


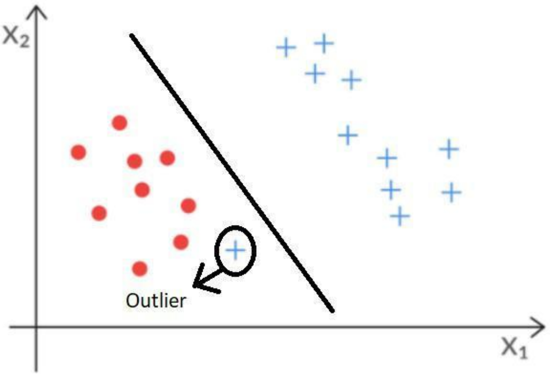
*Figure: SEQ Figure * ARABIC 7: Function of two variables*

Formula for number of variables n f= 4n+1
Extension of Boundary Value Analysis: Robustness[14].

In Robustness BVA we have two more addition to above Min– and Max+
• Min- range-------------Just below Minimum
• Max+ range -----------Just above Maximum

$$f = 6n + 1$$

On considering the range values of max and min with Extended BVA optimum cost of misclassification can be considered.



Proposed Optimum Cost of Misclassification C:

Outlier (n) = 1

Apply Extended BVA

C= 6n + 1

C = 6 *1 + 1

C = 7

V. CONCLUSION

Proposed novel idea of applying Extended Boundary Value analysis helps in solving the problem of identifying Optimum Cost of misclassification (C) in Support Vector Machine.

If C is high, a higher number of points are allowed to be misclassified or violate the margin. In this case the model is flexible, more generalizable and less likely to over fit. In other words it has high bias.

On the other hand if C is low, a lesser number of Points are allowed to be misclassified or violate the margin. In this case the model is less flexible, less generalizable and more likely to over fit. In other words it has a high variance.

REFERENCES

1. Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems Amani Yahyaoui1, Nejat Yumuşak Biomedical Research 2018 , 29 (7): 1474-1480.
2. An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data ,Ying Zhang,1 Qingchun Deng,2 Wenbin Liang,3 and Xianchun Zou BioMed Research International , 2018.
3. Data mining in healthcare–a review. Procedia,Computer Science , NeeshaJothi,WahidahHusain,etal., 72:306–313, 2015.
4. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology, Divya Tomar and Sonali Agarwal ,5(5):241–266, 2013.
5. Greg JF. Household-contact investigation for detection of tuberculosis in Vietnam. N Engl J Med 2018; 378: 221-229.
6. Sentiment Analysis using SVM: A Systematic Literature Review , Munir Ahmad1, Shabib Aftab2, Muhammad Salman Bashir3, Noureen Hameed4 , IJACSA, Vol. 9, No. 2, 2018.
7. N. T. Liu and J. Salinas, “Machine learning in burn care and research: A systematic review of the literature,” Burns, vol. 41, no. 8, pp. 1636–1641, 2015.
8. Maria Virvou, Efthimios Alepis, George A. Tsihrintzis, Lakhmi C. Jain:Machine Learning Paradigms-Advances in Learning Analytics Intelligent Systems Reference Library 158, Springer 2019, ISBN 978-3-030-13742-7.
9. Empirical Approach to Machine Learning. Studies in Computational Intelligence 800, Springer 2019, ISBN 978-3-030-02383-6.
10. Jun Jiang, Zhe Wen, Mingxin Zhao, Yifan Bie, Chen Li, Mingang Tan, Chaohai Zhang:Series Arc Detection and Complex Load Recognition Based on Principal Component Analysis and Support Vector Machine. IEEE Access 7: 47221-47229, 2019.
11. HushengGuo, Wenjian Wang:Granular support vector machine: a review. Artif. Intell. Rev. 51(1): 19-32 (2019).
12. Seokho Kang, Dongil Kim, Sungzoon Cho:Approximate training of one-class support vector machines using expected margin. Computers & Industrial Engineering 130: 772-778 (2019).
13. Tianyong Wu, Jian Zhang:Boundary value analysis in automatic white-box test generation. ISSRE 2015: 239-249.
14. Rolci Cipolatti, I-Shih Liu, Luiz A. Palermo, Mauro Antonio Rincon, Ricardo M. S. Rosa:A boundary value problem arising from nonlinear viscoelasticity: Mathematical analysis and numerical simulations. Applied Mathematics and Computation 335: 237-247 (2018).