CLASSIFICATION OF AUTHENTIC DATA USING SUPPORT VECTOR MACHINE IN AI

Aman Mishra[1*], Dr. C.L.P. Gupta[2], Deepak Kumar Singh[3]

Abstract

As we know that security is a concerning topic for each country. There are numerous intrusions launched by numerous attackers, and each attacker attempts to breach the security system in order to gain advantages in various scenarios. We can use a behavioral approach to capture the attack and classifying the type of attack becomes more important in this situation. There are so many choices of algorithms that are available, the researchers select appropriate tools to classify the type of data.

IDS (Intrusion detection system) is a software application that monitors malicious activitiesand produces reports. There are so many types of IDS systems available nowadays; some intrusion detection systems are network-based, signature-based, or anomaly-based. Many researchers work with anomalous data to classify attacks. We are using an SVM (Support Vector Machine) classifier for this purpose. The main thing in SVM is theselection of kernel, which has the primary role of classifying data classes. For this purpose, we used the RBF function as a kernel for better results.

In this paper, we are working on an AI-based classifier called SVM to classify the anomalous data. We used different types of kernels that fits with our dataset.
The information is summarized in .CSV data file this is a large dataset with a large amount of authentication-related data.

Keywords: IDS, RBF, LSVM, NSVM, kernel.

[1*]Computer Scienceand Engineering, Bansal Institute of Engineering and Technology,Lucknow
Email: kmishra446@gmail.com
[2]Computer Science and Engineering, Bansal Institute of Engineering and Technology, Lucknow
Email: hod@bansaliet.in
[3]Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur
Email: deepakk005@gmail.com

***Corresponding** Author: Aman Mishra
*Computer Scienceand Engineering, Bansal Institute of Engineering and Technology,Lucknow
Email: kmishra446@gmail.com

DOI: 10.48047/ecb/2023.12.si10.0031

I.INTRODUCTION

Vapnik introduced the Support Vector Machine (SVM), which sparked a wide interest in machine learning research. SVMs have been shown in multiple recent studies to beat alternative data categorization methodologies in terms of overall accuracy. SVMs have been used to handle problems such as text categorization, Recognition of tone and handwriting, object recognition and visual classification, processing of micro-array genomic data, and data classification. The cost parameter and kernel parameters have a considerable impact on the efficacy of SVM for certain datasets. As a result,while finding the right parameter value, the user is usually required to undertake extensive cross-validation. This is generally known as "model selection."

Time consumption is one of the practical issues in Model selection. There are various application-Many revised techniques are commonly used to minimize the time-space complexity of SVMs. Another goal is to promote the SVM algorithm block's efficiency [1]. To avoid the challenges of quadratic programming, other approaches such as the central support vector machine algorithm, the scale approach, and the Lagrangian SVM are utilized.

Data mining is becoming a more useful approach for converting raw data into useful information [3]. SVM is used for Linear and Non-Linear classification. Support Vector Machines (SVM) is a sophisticated, cutting-edge algorithm based on the Vapnik- Chervonenkis idea. The regularization abilities of SVM are excellent [4].

The term "regularization" refers to the model's generalization to new data. SVMs was created specifically to handle supervised learning classification tasks. When learning under supervision, we are given a "training dataset" for which we already have the categories.

related aspects of the SVM algorithm that might affect the result. The kernel functions utilized are included in these parameters, the SD of Gaussian kernel proportional weights for the slack variables to take into for the non-uniform distribution of labelled data, and the number of training samples. For example, we used four separate application data sets, diabetes data, heart data, and satellite data, each with its own set of characteristics, classes, number of testing and training data.

SVM has demonstrated exceptional abilities in a wide range of application notably in classification problems [2]. Its major purpose is to maximize the hyperplane, and its critical design notion is to maximize classification limits. Nonlinear separable problems should be transformed to linearly separable problemsas most challenging.

Support Vector Machine (SVM) is becoming a key tool in the development of commodities with societal repercussions, thanks to its expanding use in many applications of data mining. Identification of particles, recognition of faces, text categorization, bioinformatics, civil and electrical engineering,and other domains have all used SVMs well.

In addition to providing a brief overview of SVMs relevant to data mining issues, this research also aims to provide a review of the state of the art for the use of SVM data mining methods [5].

Kernel selection is the critical step with SVM. The picture depicts the SVM boundary conditions for opting the possible an optimal hyper-plane in which data points are separated class-wise

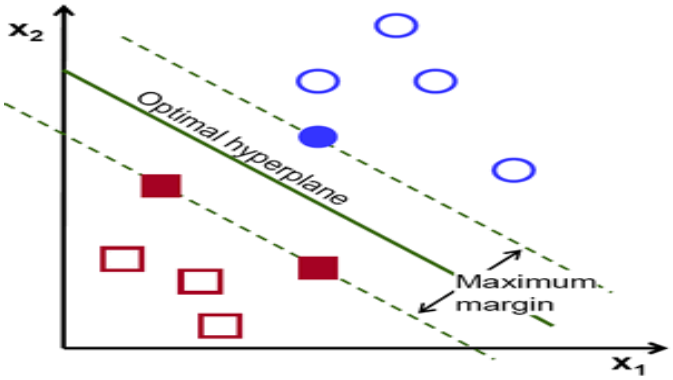

Fig. 1. Noel Bambrik. (2019). Hyper-plane with two different classes.

Retrieved from https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c.

A. Kernel-Trick

- As per the knowledge, SVM perform operation by placing dot products in between and it offers delayed solutions and the side-step of new dimensions. Consider the following:
- Allow the much-needed new space:

$$z = x^2 + y^2$$

- Specifying how the derivative will appear in provided space:

$$a \cdot b = x_a \cdot x_b + y_a \cdot y_b + z_a \cdot z_b = x_a \cdot x_b + y_a \cdot y_b + (x_a^2 + y_a^2) \cdot (x_b^2 +)$$

Mostly in kernel-trick, As a part of our attempt to incorporate the new dot product, we ask SVM classifier to complete a number of tasks. There can be a non-linear border between classes in this feature-space due to its ability to accommodate non-linear bordering between classes.

B. Support Vector Machine

A variety of classification and forecasting applications have proven to benefit from the use of SVMs [6]. SVMs were created byVladimir Vapnik in the domains of statistical inference and structural reduced risk. The generalization principle is based onthe Theory of Structural Risk Minimization (SRM), i.e., the method relies on statistical learning theory's given risk limits, enabling SVMs to capture very high dimension spaces [7].

The weights of MLP (Multilayer perceptron) classifiers are modified throughout the training phase to reduce the sum of all errors between the goal output and the network outputs. Because it is difficult to resolve decision boundaries across classes during training, and generalization ability is reliant on the training procedure, the network's performance falls dramatically for small data quantities. Decision boundaries in SVM, on the other hand, arederived straight from the training batch of data, and In feature space, separation margins can be maximized.

A space-spanning maximum fringe hyperplane that uses non- linear boundaries to categories data is referred to as an SVM. Boundaries that are not linearly formed by determining a collection of separating hyperplanes multiple data classes of points. An SVM-based hyper-plane that maximizes the separation between the margin and the hyperplane separates the positive or negative training sample from a collection of assigned training samples labelled either positive or negative. Ifno hyperplanes can split the favorable or unfavorable samples, an SVM picks a hyperplane which separates the sample as precisely as possible while retaining the gap between hyperplanes constant as little as possible.

The approach reinforces the ideal separation surface or the hyperplane that is equidistant between two classes, as per the geometrical interpretation of support vector classification (SVC). SVC is being shown once for the linearly separable case. Following that, kernel functions are sused to generate non-linear decision surfaces. Finally, slack variables are used to account for false - negative through noisy data once absolute separation of the two lessons is not desired. [8].

II. RELATED WORK

In this article, many studies have been published in the past fewyears on various SVM processes. We have some discussion of the literature on SVM.

For pattern recognition, Zhen Yang et al. presented a DE-SVC.As efficient data processing methods, Differential evolution and support vector machines can both be used.

The updated DE-SVC prediction model requires far fewer iteration steps and training time than the genetic algorithm.

They argue that SVC has more capacity for adaptive learning and predictability.

For visual generalization, Chih-Fong Tsai introduced a two- stacked generalization strategy classification, consisting of three generalizers that use the colour texture of support vector machines (SVMs). He analyzed the classification performances, hyperplane margins, and SVM support vector numbers of two training approaches for the proposed classification scheme, based on two rounds of cross-validation and no cross-validation with more accurate categorization rates,a broader hyperplane margin, and fewer support vectors, the non-cross-validation training technique excels the cross-validation training method [9].

From a historical perspective, Nahla Barakat et al. looked at theSVM research topic and rationally organized and assessed many solutions. They separated rule extraction into two groups, one dependent on the SVM (model) components used, and the other on the rule extraction technique. The algorithms' main features are then contrasted, along with their relative performance as measured by various criteria. The study finisheswith a list of potential research objectives, which suggests that rule extraction strategies are necessary for SVM

incremental and active learning, as well as additional SVM application areas.

By constructing only two information granules top-down, Yuchun Tang et al. developed granular support vector machines, a ground-breaking learning approach for problems with data categorization. The results of their Experiments on Three tasks for medical binary classification show that the proposed granular support vector machines perform as expected, the technique is an exciting approach for solving difficult classification and regression difficulties that are ubiquitous in medical and biological data processing capabilities.

Himanshu Rai and colleagues developed an innovative and effective approach for extracting and detecting iris features. They claim to have increased efficiency by using alternative edge detection algorithms for SVM and Hamming distance- based classifiers. Claim to have demonstrated that the proposed method performs extraordinarily well in terms of FAR (False acceptance rate) and FRR (False rejection rate) for both the CASIA and Check image datasets.

The rough set is a unique mathematical technique for coping with non-integrality and ambiguous information. It can evaluate and deal with a variety of confusing, contradictory and lacking data, extracting connotative information and disclosing its underlying rules. In 1982, a mathematician from Poland Z.

Recently, the applicability of the rough set theory is often emphasized in data mining and artificial intelligence.
Jui-Sheng Chou et al. used QUEST, CHAID, C5.0, CART, and GASVM, He introduced a series of classifiers that may be used to forecast the likelihood of a disagreement (a composite strategy). The two most effective classification and regression- based approaches for detecting project challenges in terms of precision are GASVM (89.30%) and C5.0 (83.25%). GASVM outperforms the reference models (C5.0, CHAID, CART, and QUEST) as well as earlier work done by 5.89–12.95 per cent.

III.ANALYTICAL LIMITATIONS, SUGGESTIONS & DISCUSSIONS

In light of the author's background, this literature review examines the uses of SVM in a range of domains, application enthusiasm and competence understanding in a certain area. For various uses, several writers have been used repeatedly. The Paper discusses pattern classifications, nuclear component classification, and the SVM technique., medical engineering, classification issues, science, applications for forecasting, The information was given by Elsevier, IEEE Xplore, Springer Link, Taylor Francis, and Inderscience.

SVM's popularity is growing by the day, as an outcome, numerous research applications based on SVM must be released in order to aid the extension of SVM's scope, both academically and practically. However, numerous studies have identified obvious disadvantages of SVM that must be addressed, like as: (1) The selection of a kernel for a certain circumstance (2) The machine's speed of operation during training and testing, (3) During the testing phase, its speed of convergence is slower, (4) The selection of elevated kernel parameters, (5) The model's implementation necessitates a large amount of memory space[10]. (6) Choosing between a parametric and a non-parametric approach to implementation. This merging of techniques and cross-disciplinary knowledge would result in new insights for SVM problem-solving.

This study looks at a few key SVM applications, but further research in domains including social, statistical, and behavioral science is anticipated. In addition, we will address the qualitative and quantitative aspects of the SVM model in our future research [11].

IV.PROPOSED WORK

A. Data Set Information:

Data were extracted from images of the authenticity and forgery of specimens that look like banknotes. For digitalization, a commercial camera frequently used for print inspection was employed. Finally, AI can used to create visualization of data and mathematical equation allowing users to see a problem in a new light. By utilizing AI in mathematics, users can increase their understanding of complex problems and gain insight into their data that would not be accessible without the assistance of AI.

B. Attribute Data:

1.Wavelet Transformed Image Variance (continuous)
2.Wavelet skewness Image transformation (continuous)
3.Wavelet kurtosis' Image transformation (continuous)
4. Image entropy (continuous)
5. class (integer)

The dataset was obtained from kaggle, which has over 50000 public datasets.

C. Application/Tools Used:

For result analysis we used python as a language and Anaconda distribution for the coding purpose. We also used different python library with spider tool that is freely available over internet.

D. Steps in SVM include:

1] During data preprocessing stage, we talked the test set's total result as a whole.
2] We then fitted the classifier to the training set.
3] The SVM classifier able to fitted upon the training data set.
4] After the creation of confusion matrix, we will now contrast the effectiveness of SVM and Logistic Regression Classifiers.
5] A confusion matrix function must be integrated.
6] Visualizing the outcome of the training set.
7] Envisioning the results of the test set
8] Finally compare the results, thereafter.

After apply our training dataset we have tested it by using different kernels on our SVM model. Every time measured its accuracy and we find that Redial Basis Function (RBF) performs very well and sigmoid kernel performs very low.

V. PERFORMANCE ANALYSIS

We produce the confusion matrix after fitting the training datasets to SVM using many of the kernel functions available in SVM. we find four are best suitable for our model. We have applied different kernels and performance results are shown below-

	precision	recall	f1-score	support
0	0.99	0.98	0.99	154
1	0.98	0.99	0.98	121
accuracy			0.99	275
macro avg	0.98	0.99	0.99	275
weighted avg	0.99	0.99	0.99	275

Figure 1: Result comes through Linear kernel

	precision	recall	f1-score	support
0	0.72	0.72	0.72	154
1	0.64	0.64	0.64	121
accuracy			0.69	275
macro avg	0.68	0.68	0.68	275
weighted avg	0.69	0.69	0.69	275

Figure 2: Result comes through Sigmoid kernel

	precision	recall	f1-score	support
0	1.00	1.00	1.00	154
1	1.00	1.00	1.00	121
accuracy			1.00	275
macro avg	1.00	1.00	1.00	275
weighted avg	1.00	1.00	1.00	275

Figure 3 : Result comes through RBF kernel

	precision	recall	f1-score	support
0	1.00	0.92	0.96	154
1	0.90	1.00	0.95	121
accuracy			0.95	275
macro avg	0.95	0.96	0.95	275
weighted avg	0.96	0.95	0.95	275

Figure: Result comes through Polynomial kernel

Fig. 2. Performance metric using four different kernels

After applying four different kernel functions like linear kernel, Sigmoid kernel, redial basis kernel (RBF), and Polynomial kernel, we found that each kernel

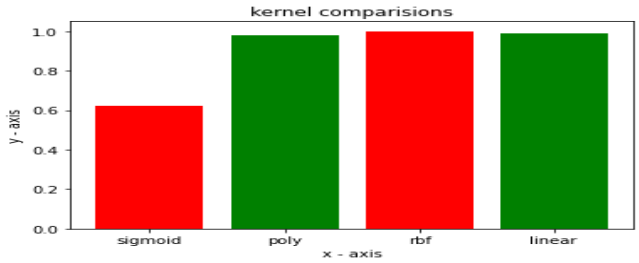

Fig. 3. Comparisons of result using different kernel functions

In the above comparison diagram, we can analyze that the sigmoid kernel is not fitted to our dataset but other are performing well, but RBF kernel is performing best among all, we can clearly see this in the below diagram.

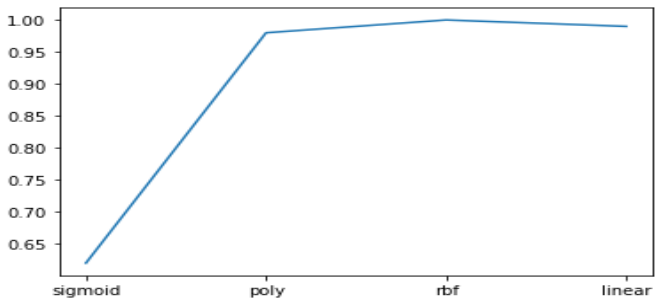

Fig. 4. Other comparisons of result using different kernels

We can see that the RBF function performing very well and giving 100% accuracy to classify our dataset. The next table is the conclusion table showing the accuracy of SVM based on different kernels.

Table 1. Showing the accuracy based on different kernel

Kernel	Accuracy
RBF	100%
Linear	99%
Polynomial	95%
Sigmoid	69%

VI.CONCLUSION

A SVM solution's blinding whiteness, which implies that most of the entries in the solution-vectors become zero, is a desirable quality. Therefore, an implementation is done with SVM produced a high degree of precision with such a large database that has already been bound mostly by computation time used for optimization and still more help can be provided by least - square Classifier, which resolves linear equations instead of QP issues. Additionally, we worked to improve the SVM by utilizing a variety of kernel techniques, such as Gaussian, Polynomial, Linear, and Redial, and have achieved very good accuracy up to 100% by using the RBF (Radial basis function) kernel.

VII. REFERENCES

1. Nayak, Janmenjoy, Bighnaraj Naik, and H. S. Behera. "A comprehensive survey on support vector machine in data mining tasks: applications & challenges." International Journal of Database Theory and Application 8.1 (2015): 169-186.
2. Cervantes, Jair, et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends." Neurocomputing 408 (2020): 189-215.
3. Hasseim, Anith Adibah, Rubita Sudirman, and Puspa Inayat Khalid. "Handwriting classification based on support vector machine with cross validation." Engineering 5.5 (2013): 84-87.
4. Salcedo-Sanz, Sancho, et al. "Support vector machines in engineering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4.3 (2014): 234-267.
5. Lumbanraja, Favorisen R., et al. "Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science Journal)." Journal of Physics: Conference Series. Vol. 1751. No. 1. IOP Publishing, 2021.
6. Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends." Knowledge-Based Systems 226 (2021): 107134.
7. Durgesh, K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." Journal of theoretical and applied information technology 12.1 (2010): 1-7.
8. Amarappa, S., and S. V. Sathyanarayana. "Data classification using Support vector Machine (SVM), a simplified approach." Int. J. Electron. Comput. Sci. Eng 3 (2014): 435-445.
9. Sarkar, Anurag, et al. "Text classification using support vector machine." International Journal of Engineering Science Invention 4.11 (2015): 33-37.
10. Wei, Liwei, Bo Wei, and Bin Wang. "Text classification using support vector machine with mixture of kernel." Journal of Software Engineering and Applications 5 (2012): 55.
11. "The Sixth International Symposium on Neural Networks (ISNN 2009)", Springer Science and Business Media LLC, 2009