A Nonlinear Support Vector Machine Analysis Using Kernel Functions for Nature and Medicine

Benajiba Yassin[1,*], *Chrayah Mohamed* [2], *Al-Amrani Yassine*[3]

[1] TIMS Laboratory, FS Tétouan, Abdelmalek Essaadi University, Morocco.
[2] TIMS Laboratory, ENSA Tétouan, Abdelmalek Essaadi University, Morocco.
[3] TIMS Laboratory. FP Larache, Abdelmalek Essaadi University, Morocco.

Abstract. After the emergence of Artificial Intelligence (AI), great developments have taken place in the fields of science, economics, medicine and all other fields that use computer science. Along with the resulting developments in these fields, artificial intelligence has also solved many intractable problems, such as predicting specific serious diseases, determining future product sales, as well as analyzing and studying big data in the shortest possible time ... SVM is one of the most important technologies in this field of artificial intelligence that goes into supervised methods, and which every machine learning expert should have in his/her arena. For this reason, in this article, we studied this technique and determined its advantages and disadvantages as well as its fields of application. Next, we applied this technique to three different databases, using four basis change functions, and we compared the results obtained to determine the best way to use the basis change functions.

Keywords : AI · SVM · KERNEL FUNCTION.

1 INTRODUCTION :

The idea of artificial intelligence first appeared in 1956 during a conference at Dartmouth College. Researchers have since sought to find a way to simulate human intelligence with a machine, but they were not successful until in the early 1980s, when a program called "Expert Systems" was born. This program simulates human intelligence, human knowledge and analytical skills using synthetic skills. Since then, artificial intelligence has been developing, advancing and being used in many fields.

Artificial intelligence is a major scientific discipline that studies the methods and techniques of solving logical problems, and creating algorithms. There are four types of artificial intelligence: a) interaction machines, b) limited memory, c) theory of the mind and d) self-awareness. SVM is one of the most important artificial intelligence technologies and it is classified as one of the techniques under supervision, and it solves many problems of data classification and analysis.

2 SUPPORT VECTOR MACHINE :

The Support Vector Machine (SVM) is an important, simple, supervised learning algorithm that every programmer and machine learning expert should have.in their arena. It can be used for regression and classification tasks. However, it is used mostly for classification purposes, and the goal of the Support Vector Machine algorithm is to find a hyperplane in an N-dimensional space that clearly separate data points.

For a more precise understanding of how this method works, we will translate it mathematically:

Suppose A is a set of n data / class pairs, defined by:

$$\mathbf{A} = \{(x_1 , y_1), (x_2 , y_2), \dots , (x_n , y_n)\}$$

Such that: $y_i \in \{-1.1\}$ is the label indicating whether or not an observation belongs to a class, the number n designates the dimension of the database, and xi the ith sample of this base. Any sample xi has p descriptive variables, i.e. it can be expressed as follows:

$$x_1 = \begin{pmatrix} x_{i1}\ x_{i2}\ x_{i3} \\ \vdots\ x_{ip} \end{pmatrix}$$

In practice, to be able to apply the SVM method correctly, it is first necessary to choose the descriptive parameters well because the selection of these parameters is crucial in the classification of the data.

The second step is to find the optimal hyperplane which will divide the training data in half so that all points of the same type are on the same side of the hyperplane because the plane will be divided into two different parts, and each part will have the same type of points.

* Corresponding author: yassin.benajiba@etu.uae.ac.ma

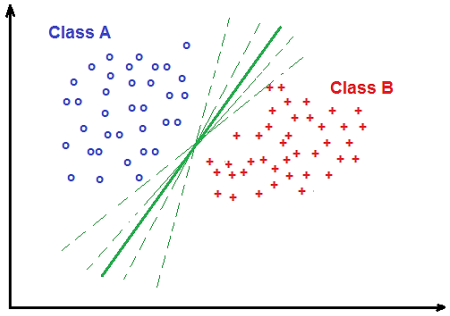

Fig. 1. The infinity of the hyperplane of separation.

Then, we seek the optimal hyperplane to separate these two types of points, i.e. we seek to maximize the distances between the points of the learning classes and the hyperplane, this distance is called the margin, and the minimum distance points are called support vectors. There are two types of separation methods: linear separation and non-linear separation.

2.1 Linear separation or linear SVMs:

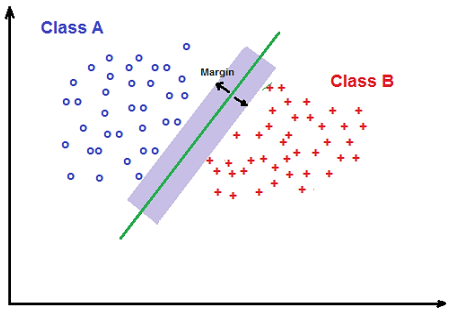

Fig. 2. The optimal hyperplane and the maximum margin.

Linearly separable cases are the simplest cases of SVM, because they make it easy to find the hyperplane (line) separating the classes, so we are just trying to maximize the classifier margin, to find a good separating hyperplane.

2.2 Nonlinear separation or nonlinear SVMs:

In real SVM applications, classes cannot be separated linearly, so we are working with nonlinear SVM to work around this problem, That is to say, by applying a nonlinear transformation to the data to change dimension and easily find a hyperplane classification in this new space, and also to give the classifier more freedom to correctly classify the points even if they are initially points on the wrong side of the initial hyperplane (non-separable categories).

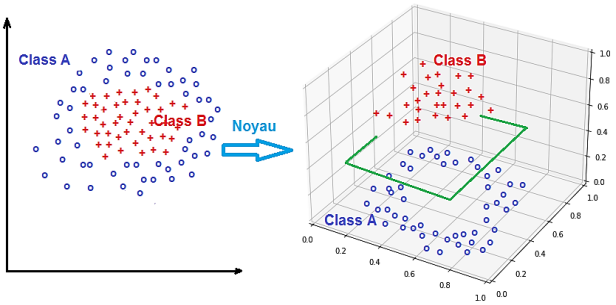

Fig. 3. Example of a nonlinearly separable problem.

The optimization problem is written as follows:

$$\{min\ \frac{1}{2}\|w\|^2 + C\ \sum_{i=1}^{n}\ \varepsilon_i\ \forall i,\ y_i(w.x_i + b)\ \geq\ 1\ -\ \varepsilon_i$$

Where w and b are parameters of the hyperplane, C is the weight given to samples located on the wrong side of the separation boundary (also called regularization constraint), εi are parameters which allow to consider badly classified points.

Thanks to the kernel trick, the dual problem will be:

$$\{max_{\alpha}\ \sum_{i=1}^{\alpha_i}\ \alpha_i\ - \frac{1}{2}\sum_{i,j=1}^{n}\ \alpha_i\alpha_j y_i y_j\ K(x_i, x_j)\ \forall i,\ \ 0\ \leq \alpha_i\ \leq\ C\ \ \sum_{i=1}^{n}\ \alpha_i y_i\ = 0$$

Where α_i are the Lagrange multipliers, and K (. , .) Represents the kernel function.

In practice, we find that there are few families of kernel functions still in use, and to choose the right type, the user has to perform several tests to determine which is the best for his application. In our applications we use the following kernel functions:

- **Linear :** $\mathbf{K}(x_i; x_j) = x_i^T . x_j$
- **Polynomiale :** $\mathbf{K}(x_i; x_j) = (x_i^T . x_j + 1)^d$
- **Gaussien :** $\mathbf{K}(x_i; x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2})$
- **Sigmoïde :** $\mathbf{K}(x_i; x_j) = \tanh(\mathbf{k}\ x_i^T . x_j +$)

The optimal solution to this problem is therefore to determine the result of this function necessary for the classification of each sample:

$$f(x) = sign(\sum_{i=1}^{M}\ \alpha_i y_i K(x_i, x)\ +\ b)$$

Where x_i and y_i are respectively the support vectors and their membership classes.

2.3 The advantages of SVM:

• SVM is very efficient if the number of dimensions is high.

• If the dimension of the space is greater than the number of training samples, then SVM becomes efficient. For the decision, SVM only uses a few training samples

(the support vectors). As a result, it does not use much memory.

2.4 The disadvantages of SVM:

• SVM performance is poor if the number of attributes is much greater than the number of samples.

• Since SVMs are methods of discrimination between classes, they do not provide probability estimates.

3 THE IMPORTANCE OF SVM IN NATURE AND MEDICINE:

3.1 Breast cancer:

Breast cancer is an important and common disease that negatively affects the health of women and most often diagnosed in women worldwide, it accounts for 11.6% of the total number of cancer incidence. It is the leading cause of cancer death in women around the world. Research shows that experienced doctors can detect cancer with an accuracy of 79% [1], while an accuracy of 91% (sometimes up to 97%) can be achieved using SVM, which shows the importance of the use of SVM in the diagnosis of breast cancer [2,3,4,5].

3.2 Superb Iris flower:

Iris is a beautiful and elegant plant that is preferred by women. Irises are mostly afraid of the sun because it prevents them from flowering properly, and they are afraid of watery soils because they like semi-dry soil, so finding this plant is a bit tricky, and difficult to identify from a distance among the rest of the plants. SVM can be used to find the iris flower and distinguish it among the rest of the plants in a simple and fast way [6],

4 APPLICATION AND DISCUSSIONS:

4.1 Descriptions of the data used:

Premier Data: **ticket authentication** [1]

Table 1. The first five rows of ticket authentication data set.

Variance	Skewness	Curtosis	Entropy	Class
3.6216	8.6661	- 2.8073	-0.44699	0
4.5459	8.1674	-2.4586	-1.4621	0
3.866	-2.6383	1.9242	0.10645	0
3.4566	9.5228	-4.0112	-3.5944	0
0.32924	-4.4552	4.5718	-0.9888	0

Data Set Information:

These data were obtained from images taken from real samples of banknotes and counterfeit samples that resemble banknotes. It is made up of 400 x 400 pixels.

Attribute Information:

1. variance of Transformed image.
2. skewness of Transformed image.
3. curtosis of Transformed image.
4. entropy of image.
5. Class.

Secondary data: Iris datasets [2]

Table 2. The first five rows of of Iris datasets.

Id	Sepal LengthCm	Sepal WidthCm	PetalLength Cm	Petal WidthCm	Species
1	5,1	3,5	1,4	0,2	Irissetosa
2	4,9	3,0	1,4	0,2	Irissetosa
3	4,7	3,2	1,3	0,2	Irissetosa
4	4,6	3,1	1,5	0,2	Irissetosa
5	5,0	3,6	1,4	0,2	Irissetosa

Data Set Information:

This dataset contains 3 species and each type contains 50 different cases, with each case referring to a type of iris plant.

Attribute Information:

1. id
2. sepal Length in cm
3. Petal Width in cm
4. petal Length in cm
4. petal width in cm
5. Species :

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Third Data: Wisconsin Breast Cancer Data Set (Diagnosis) [3]

Table 3. The first five rows of of Breast Cancer Data Set.

Id	Diagnosis	Radius mean	Texture mean	…	Symmetry worst	Fractal dimension worst
842302	M	17,99	10,38	…	0,4601	0,1189
842517	M	20,57	17,77	⋯	0,275	0.08902
8430090	M	19,69	21,25	⋯	0,3613	0,08758
8434830	M	11,42	20,38	⋯	0,6638	0,173
8435840	M	20,29	14,34	⋯	0,2364	0,07678

Data Set Information:

Features are calculated from a fine needle aspiration (FNA) scanned image of a breast mass. They describe the characteristics of the cell nuclei present in the image.

Attribute Information:

1.Identification number.
2.Diagnosis (M = malignant, B = benign)

3-32. Ten real-value characteristics are calculated for each cell nucleus:

a) Radius (average of the distances from the center to the points of the perimeter).
b) Texture (standard deviation of gray-scale values)
c) Perimeter.
d) Area.
e) Fineness (local variation of radius lengths).
f) Compactness (perimeter^2 / area -1.0).
g) Concavity (severity of concave portions of the contour).
h) Concave points (number of concave parts of the contour).
i) Symmetry.
j) Fractal dimension (coastline approximation -1)

The mean, standard error, and "worst" or more (average of the three largest values) of these characteristics were calculated for each image, resulting in 30 characteristics.

4.2 Results and discussion:

When applying the SVM algorithm on the three available datasets, we recorded all the results obtained for each data and for each change of kernel function, and we obtained the following results:

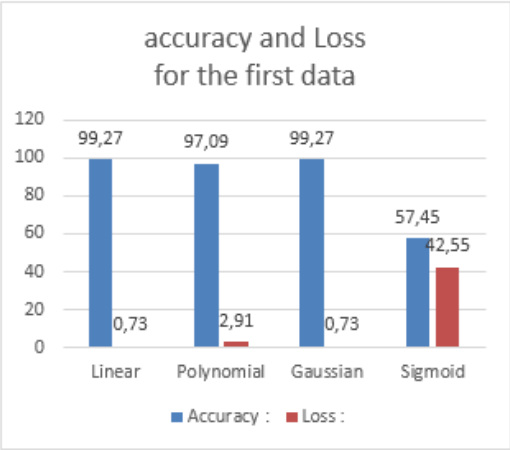

Fig. 4. The curve of variation of the Accuracy and the Loss for the first data

We noticed that the best precision we get in the first data is when we use the linear kernel function or the Gaussian function, because when we use both we get an accuracy equal to 99.27%.

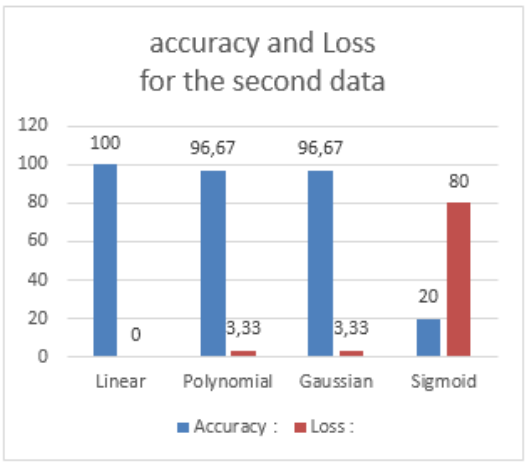

Fig. 5. The curve of variation of the Accuracy and the Loss for the second data

For the second data, the best accuracy we get is when using the linear function, because for the first group the accuracy was 100%, which is the best classification accuracy we got when the application of all the functions to the three Data.

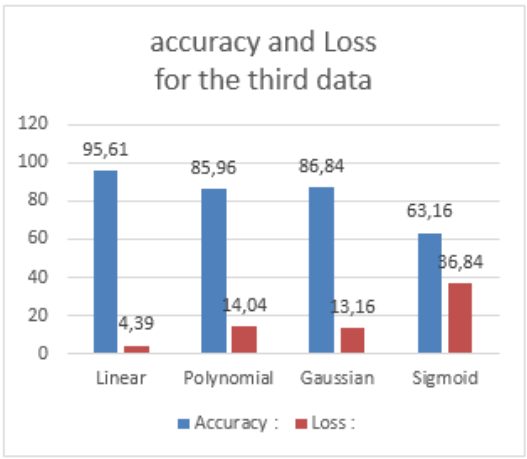

Fig. 6. The curve of variation of the Accuracy and the Loss for the third data

As for the third database, the highest accuracy rate is 95.61%, is reached when using the linear function.

We also note that the worst accuracy we get in all of this data is when using the " Sigmoid " kernel function. This does not mean it is always inefficient, as you may find it to be much better than other functions in other cases, especially when the number of classes becomes very large, in which case the linear function becomes less precise.

And here are the confusion matrices we found for each app:

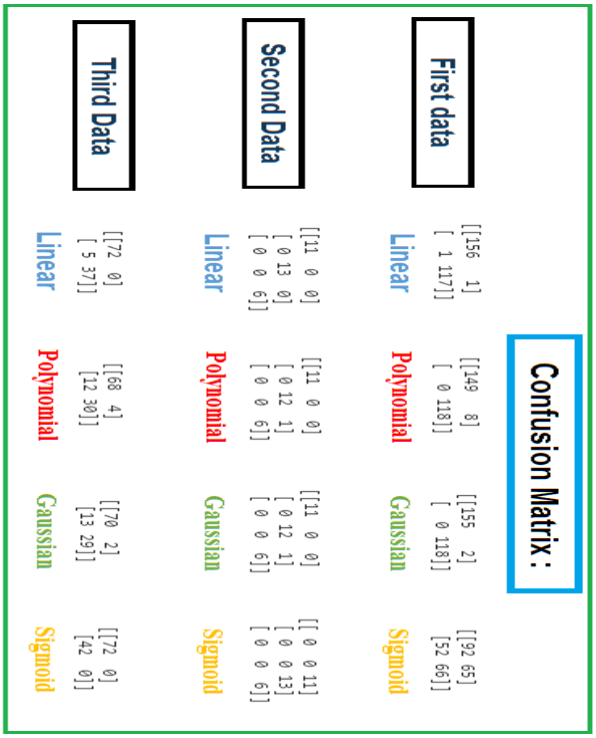

Fig. 7. The matrix of confusion.

Regarding the confusion matrices, we also noticed that the best confusion matrix we got was when we used the linear kernel function in the second data, where all the data was laid out in the right place. we also noticed that the Sigmoid kernel function is misclassifying data as it classifies all this data into one category, for the second data as well as for the third data, which also shows that they cannot be used to classify these data types.

5 CONCLUSION:

SVM is one of the most important machine learning algorithms for data classification because its accuracy can reach 100% as we have seen before. It is also a good classifier when there are many classes and its use has become widespread in most fields and sciences. And Often, nonlinear SVMs are used in classification using the use of kernel function, and to get the best precision, the user should try all the functions and then compare the results to get the best function to use.

References

[1]. G.Hamed, M.Abd El Rahman Marey, S.Amin, M.Fahmy Tolba, Deep Learning in Breast Cancer Detection and Classification, March 2020.

[2]. M. Divyavani, K. Govindaswamy, An Analysis On SVM & ANN Using Breast Canser Dataset, January 2021.

[3]. P.P. Golagani, T.S. Mahalakshmi, K. Beebi Shaik, Supervised Learning Breast Cancer Data Set Analysis in MATLAB Using Novel SVM Classifier, January 2021.

[4]. S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study”, 2018.

[5]. S. G. Durai, S. H. Ganesh, and A. J. Christy, “Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer,” In Computing and Communication Technologies (WCCCT), 2017 World Congress on 2018.

[6]. Z.F. Hussain, H.R. Ibraheem, M. Aljanabi, A.H. Ali, A new model for iris data set classification based on linear support vector machine parameter's optimization, February 2020.

[7]. J. Cervantesa, F. Garcia Lamont, L. Rodrguez Mazahuab, A. Lopez: A comprehensive survey on support vector machine classification: Applications, challenges and trends. 2019.

[8]. M. El Mountassir, G. Mourot, S. Yaacoubi, D. Maquin: SVM for better classification of Guided Waves monitoring data, April 2016.

[9]. V.J. Gaikwad, “Detection of Breast Cancer in Mammogram using Support Vector Machine”, International Journal of Scientific Engineering and Research, 2016.

[10]. S. Chakrasali, M. Akshata, B.V. Aparna, S. Donthi and N. Jain, “‘A Comparative Study between Contourlet and Wavelet Transform for Medical Image Registration and Fusion”, International Journal of Computer Science and Network Security, 2015.

[11]. H. Zamani HosseinAbadi, R. Amirfattahi, B. Nazari, H. R. Mirdamadi, and S. A. Atashipour, GUW-based structural damage detection using WPT statistical features and multiclass SVM, Appl. Acoust., vol. 86, pp. 5970, Dec. 2014.

[12]. M. Hassan, R. Rajkumar, D. Isa, and R. Arelhi, Defect Classification by Using NonDestructive Testing and Improved Support Vector Machine Classification, International journal of engineering and innovative technology, vol. 2, no. 7, pp. 8593, 2013.

[13]. X. Li, Structural Damage Classification using Support Vector Machine, thse de doctorat, 2012.

[14]. M. Sewak, P. Vaidya, C.C. Chan, Z.H. Duan, SVM Approach to Breast Cancer Classification, December 2007.