Classification of credits in the german market via SVM: Attempt at modeling

AZIZ CHIHAB

Department of mathematics/ University Ibn Tofail, Kenitra, Morroco

aziz.sma.1994@gmail.com

Mohammed Kaicer

Laboratory of Analysis, Geometry and Applications

mohammed.kaicer@uit.ac.ma

Keywords— Machine Learning, Deep learning, SVM, Credit market, prediction.

I. SUMMARY:

The work is divided into two main parts. The first part is the theoretical part in which we have defined basic notions of artificial intelligence, machine learning and these supervised and unsupervised learning algorithms and then the SVM algorithm, which is based precisely to integrate complexity control into the estimation; that is, the number of parameters that is associated in this case with the number of Supports vectors, and for this we chose to work in the practical part by the support algorithm for machine vectors , and we obtained better results in terms of the accuracy of the prediction which is increased up to 75%.

II. INTRODUCTION ET METHODOLOGY

A. Introduction

In the practical part of this work, we followed the following steps .First, we collected the data from the Learning UCI machine repository, and then in the second stage we did a data pre-processing (Cleaning and Encoding). Then to develop a model of machine Learning the Data Set has been divided into two parts: Train Set and Test Set. with the Train Set data we have developed a transformation function called a transformer which allows us to process us to drive an estimator. Once this step is completed, the transformer and estimator are used to transform the Test Set data to obtain a new prediction. Then we test the three models(SVM model with "linear" kernel, SVM model with "polynomial" kernel and SVM model with "RBF" kernel)

B. Methodology

In our work we used:

- support vecteur machine(SVM) algorithm .
- SVM with "linear" kernel.
- SVM with "polynomial" kernel .
- SVM with "RBF" kernel .
- k-nearest neighbor (k-NN) et k-means .
- Cross-validation technique.
- Supervised learning algorithms.
- Unsupervised learning algorithms.

III. RESULTS AND INTERPRETATIONS:

A. Numerical results

With the cross-validation technique that gives the best parameters that give the right estimator for each model the following results are obtained:

- SVM with "linear" kernel:
  svm.SVC(kernel =′ linear′,C = 0.1) .
- SVM with "polynomial" kernel:
  svm.SVC(kernel = 'poly', degree = 1).
- SVM with "RBF" kernel :
  svm.SV C(kernel =′ rbf′ ,C = 1000, gamma = 0.000001).

		Training		Testing	
Kernel	RÉSULTATS	1 : BONE	2 : MAUVAIS	1 : BONE	2 : MAUVAIS
Linéaire (C=0.1)	Précision	0.75	0.63	0.75	0.77
	Recall	0.93	0.28	0.96	0.31
	F1_score	0.83	0.39	0.84	0.44
	support	632	268	68	32
Polynômial (Degré= 1)	Précision	0.74	0.69	0.73	0.88
	Recall	0.96	0.20	0.99	0.22
	F1_score	0.84	0.31	0.84	0.35
	support	632	268	68	32
RBF (C=1000 ET gamma=0,000001)	Précision	0.73	0.75	0.70	0.75
	Recall	0.98	0.15	0.99	0.09
	F1_score	0.84	0.25	0.82	0.17
	support	632	268	68	32

Table 1: Summary of Results Achieved With Cross-Validation Technique

B. Graphic results:

`temps d exucution NOYAU linear : 198.1477530002594`

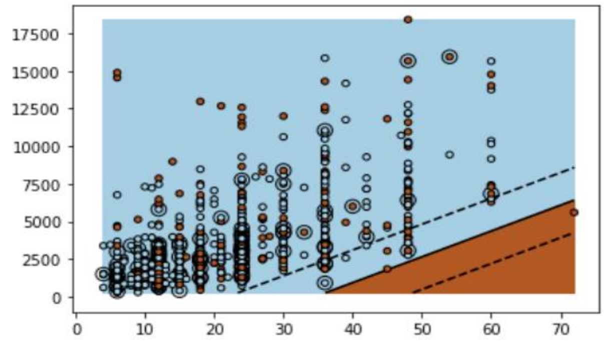

Figure 1:Graph in the case of the "linear" kernel

`temps d exucution noyau poly : 258.98491287231445`

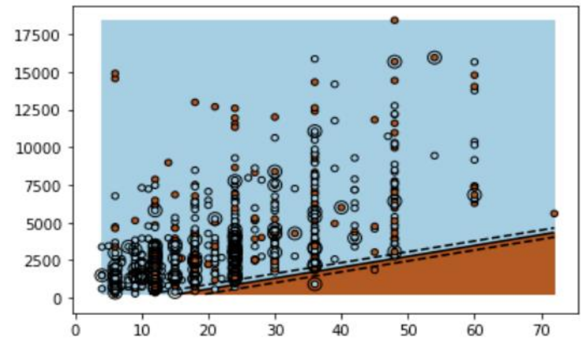

Figure 2:Graph in the case of the "polynomial"kernel

`temps d exucution noyau RBF : 12.293589353561401`

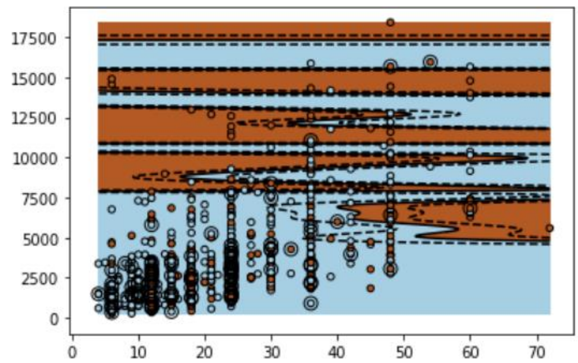

Figure 3:Graph in the case of the "RBF" kernel

This brief is an introduction to scientific research in the field of artificial intelligence and to develop models of machine learning, which being able to classify credit in the market in general one especially in the German market. This work is designed to solve non-linear problems using a non-linear kernel SVM algorithm, which allows a good visualization of the classification and data classes of the best results in terms of production accuracy, In our work we got 75% accuracy.

C. Interpretation

The linear model has greater precision in the classification of unsuccessful treatments ("1", 75%). In particular, the 96% recall suggests that almost none of the unsuccessful treatments are missing from the entire test sample. However, the model is almost unable to identify successful cases ("2"). It captures only about 31% of potential candidates and includes many false positives. Moreover, the proportion of true positive classifications that are truly positive is very, very low (45%).

The polynomial model has a higher precision in the classification of unsuccessful treatments ("1", 73%). In particular, the 99% recall suggests that almost none of the unsuccessful treatments are missing from the entire test sample. However, the model is almost unable to identify successful cases ("2"). It captures only about 22% of potential candidates and includes many false positives. Moreover, the proportion of true positive classifications that are truly positive is very, very low (35%).

The RBF model has a higher precision in the classification of unsuccessful treatments ("1", 70%). In particular, the 99% recall suggests that almost none of the unsuccessful treatments are missing from the entire test sample. However, the model is almost unable to identify successful cases ("2"). It captures only about 9% of potential candidates and includes many false positives. Moreover, the proportion of true positive classifications that are truly positive is very, very low (17%).

By comparing the three classification ratios of these three models it was concluded that the optimal model for classifying our database is the linear kernel which has an accuracy of 77% and f1-score 44% for the second class "bad credit". And among my future work perspective it has improved the performance of this modeled so as to increase the f1-score and precision.

IV. REFERENCE

[1] Nilsson, Nils J. "John McCarthy." National Academy of Sciences (2012): 1-27.
[2] Abe, Shigeo. Support vector machines for pattern classification. Vol. 2. London: Springer, 2005.
[3] McGarry, Thomas J., et al. "Classification system for partial edentulism." Journal of Prosthodontics 11.3 (2002): 181-193.
[4] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts,

tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[5] Berrani, Sid-Ahmed, Laurent Amsaleg, and Patrick Gros. "Recherche par similarités dans les bases de données multidimensionnelles: panorama des techniques d'indexation." INGENIERIE DES SYSTEMS D INFORMATION 7.5/6 (2002): 9-44.

[6] Mack, Yue-Pok. "Local properties of k-NN regression estimates." SIAM Journal on Algebraic Discrete Methods 2.3 (1981): 311-323.

[7] Zouhal, Lalla Meriem, and Thierry Denoeux. "An evidence-theoretic k-NN rule with parameter optimization." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 28.2 (1998): 263-271.

[8] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[9] Dangeti, Pratap. Statistics for machine learning. Packt Publishing Ltd, 2017.

[10] Theobald, Oliver. Machine learning for absolute beginners: a plain English introduction. Scatterplot Press, 2017.

[11] Burkov, A. "The Hundred-Page Machine Learning Book by Andriy Burkov." Expert Systems 5.2 (2019): 132-150.

[12] Azencott, Chloé-Agathe. Introduction au machine learning. Dunod, 2019.

[13] Hilali, Hassane. Application de la classification textuelle pour l'extraction des règles d'association maximales. Diss. Université du Québec à Trois-Rivières, 2009..

[14]Descôteaux, Steve. Les règles d'association maximale au service de l'interprétation des résultats de la classification. Diss. Université du Québec à Trois-Rivières, 2014.