Rapport de Projet de Machine Learning

CHELIKH Idriss, M1 SAR, 3704957

***Abstract*—Ce document fait office de compte-rendu sur le travail effectué dans le cadre d'un projet de Machine Learning. Il porte sur la reconnaissance de lettres manuscrites. Différentes méthodes de classification seront utilisées et leurs performances seront discutées.**

I. Introduction

Un des intérêts du Machine Learning est d'apporter des solutions à des problèmes de classification. Ainsi, ce projet se concentre sur la reconnaissance de caractères manuscrits à l'aide de différentes méthodes d'apprentissage supervisé. Nous discuterons des performances de chaque méthode avec pour indicateur demandé le taux de reconnaissance. Nous comparerons quatre méthodes:

- le KPPV, avec et sans ACP,
- les forêts d'arbres aléatoires,
- le SVM multi-classes,
- et les réseaux de neurones.

II. KPPV

Dans cette première section, nous discuterons des résultats obtenus par un classificateur KPPV avec ou sans ACP.

A. KPPV sans ACP

Nous nous intéressons dans un premier temps à la classification sans ACP.

Nous cherchons à déterminer quel set de paramètres permet un taux de reconnaissance optimal. 2 paramètres sont donc essentiels ici:

- La métrique utilisée pour mesurer la distance des plus proches voisins,
- Le nombre K de plus proches voisins à utiliser.

Aussi pour faciliter le choix de paramètres, nous effectuons pour chaque métrique (de Manhattan et euclidienne donc) une série d'entraînement faisant varier le nombre de plus proches voisins à chaque itération de la série. Cela permet donc de choisir la métrique et le nombre de voisin optimal pour notre problème. Les résultats de ces séries sont explicités sur les figures 1 et 2 .

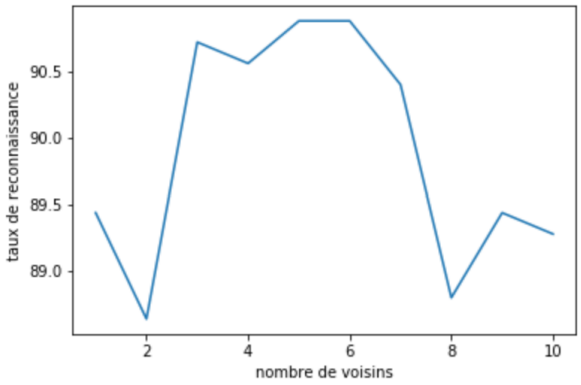

Fig. 1. Taux de reconnaissance en fonction du nombre de n voisins pour une classification en KPPV sans ACP pour métrique euclidienne

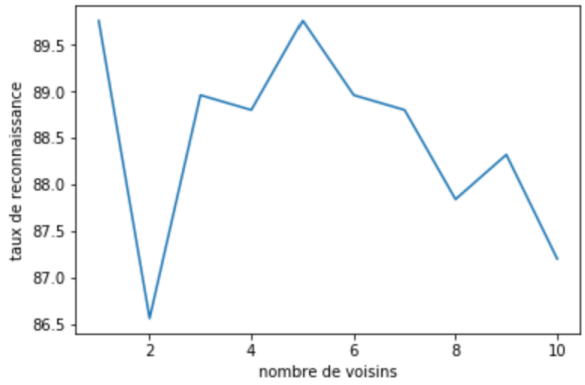

Fig. 2. Taux de reconnaissance en fonction du nombre de n voisins pour une classification en KPPV sans ACP pour métrique de Manhattan

Nous en déduisons que la métrique euclidienne est la plus adaptée pour notre problème car celle ci renvoi une valeur moyenne plus élevée que Manhattan pour ce qui est du taux de reconnaissance. Enfin, on trouve un nombre k de voisin optimal à 5

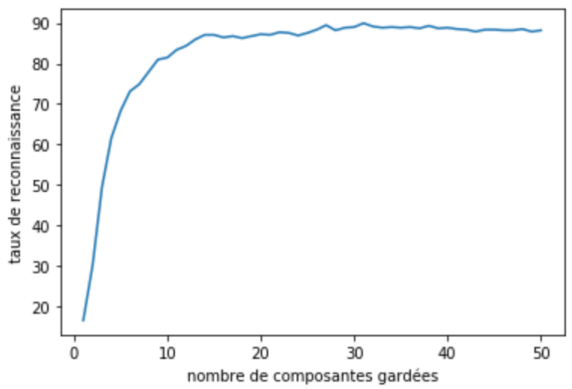

Fig. 3. Évolution du taux de reconnaissance selon la métrique Euclidienne en fonction du nombre de composantes pour une classification en KPPV avec ACP

avec un taux de reconnaissance de 90.9%.
Toutefois dans la pratique, ce classificateur demande des ressources importantes de calculs ce qui peut gêner son utilisation en temps réel pour de la saisie numérique.

B. KPPV avec ACP

Le problème de performance cité précédemment peut être pallié en ne conservant que les composantes principales des images de la base d'entraînement et de test malgré une diminution possible de la précision du modèle. Aussi nous allons maintenant étudier la classification en KPPV avec ACP. Le phénomène à étudier est la variation du taux de reconnaissance selon le nombre de composantes gardées sous ACP jusqu'à atteindre le temps d'exécution du KPPV sans ACP. La métrique utilisée sera l'euclidienne car déterminée précédemment comme adéquate pour le KPPV sans ACP.

La figure 3 représente l'évolution du taux de reconnaissance en fonction du nombre de composantes gardées (allant de 1 à 50). Nous voyons que le taux de reconnaissance se stabilise à partir de seulement 20 composantes.

Nous pouvons donc conclure que le KPPV avec ACP minimise les pertes de performances au niveau du taux de reconnaissance. Il est en effet possible d'atteindre un taux de reconnaissance de 89.9% pour un ACP à 32 composantes principales, et ce tout en ayant un temps de calcul faible.

III. Forêt d'arbres aléatoires

Pour cette section, un classificateur par forêt d'arbres aléatoires fut implémenté.
Cette classification fait intervenir 2 paramètres :

- la profondeur
- le nombre d'estimateurs utilisés par le modèle

Afin de déterminer les valeurs optimales de ces derniers, nous utilisons les outils présents sur la librairie sklearn tel que $GridSearchCV$ qui permet de retourner ces valeurs optimales en fixant un intervalle de recherche.

En faisant varier la profondeur entre 5 et 15 et le nombre d'estimateurs entre 100 et 200 avec un pas de 10, nous obtenons un taux de reconnaissance de 92.16% maximal pour une profondeur de 11 et un nombre d'estimateurs de 153.

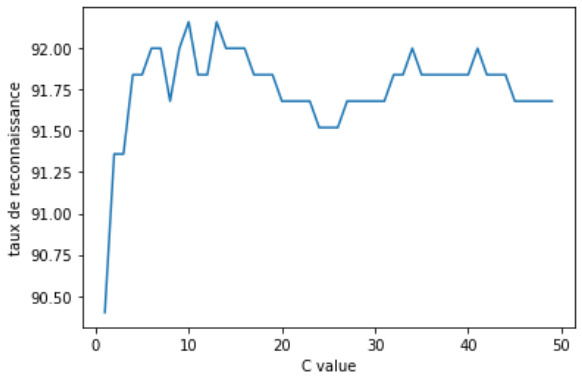

Fig. 4. Évolution du taux de reconnaissance en fonction du paramètre de régularisation pour une classification SVM avec un noyau polynomial

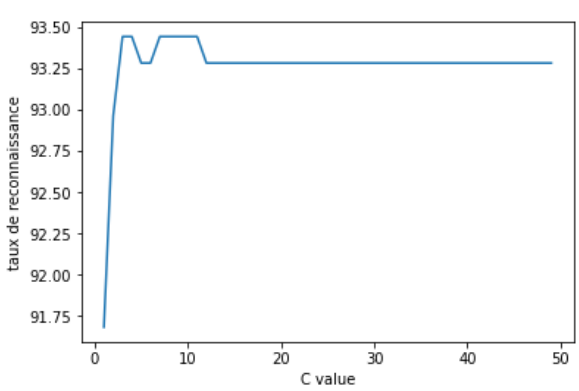

Fig. 5. Évolution du taux de reconnaissance en fonction du paramètre de régularisation pour une classification SVM avec un noyau RBF

IV. SVM MULTI-CLASSES

Dans cette section, nous implémenterons et discuterons des résultats obtenus pour une classification SVM en multi-classes.

Dans le cas d'un modèle SVM multi-classes il y a 2 paramètres principaux à prendre en compte :

- le type de noyau utilisé,
- le sigma ou C, caractéristique de la tolérance pour la classification aux frontières.

La démarche d'optimisation consiste donc à extraire les variations de la précision du modèle en fonction du C et ce pour chaque noyau étudié (ici le linéaire, polynomial et RBF (pour radial basis function).

Les figures 4, 5 représentent les taux de reconnaissance en fonction du paramètre C et pour chaque noyau étudié. Pour le noyau Linéaire, le taux de reconnaissance ne dépend pas du paramètre de régularisation, et vaut 89.5% quelque soit C. Enfin pour les 2 autres noyaux nous déduisons des courbes mentionnées, les résultats suivants :

- pour un noyau RBF un taux de reconnaissance de 93.4% pour une valeur du paramètre C de 4
- pour un noyau polynomial un taux de reconnaissance de 92.3% pour une valeur du paramètre C de 10

Il est à noter que les performances de calculs sont excellentes pour la méthode SVM et ce quelque soit le noyau choisi(calcul et entraînement avec un temps inférieur à la seconde).

De ces résultats, nous en déduisons que le classificateur SVM multiclasses avec noyau RBF est le plus précis.

V. RÉSEAU DE NEURONES

Le dernier type de méthode de classification que nous allons traiter est celle des réseaux de neurones. Pour cela un premier réseau de neurones simple fut modélisé. Puis une couche dense de neurones fut ajoutée pour finir avec un réseau convolutionnel CNN.

A. Réseau de neurones simple

Le premier réseau modélisé est composé uniquement d'une couche d'activation "softmax". Nous fixons alors un batch à 200, un taux d'apprentissage à 0.1 et un nombre d'epochs de 200. Avec la fonction de visualisation des résultats donnée en TP7, on obtient l'évolution suivante sur la figure 7.

Nous observons alors que le taux de reconnaissance pour la base de test se stabilise autour dès 50 epochs. Aussi en évaluant ce réseau simple on obtient un premier taux de reconnaissance de 88.4%.

Il nous est alors possible d'améliorer l'apprentissage du réseau de neurone en ajoutant une couche cachée de 128 neurones (même démarche que lors du TP7). Avec un réseau

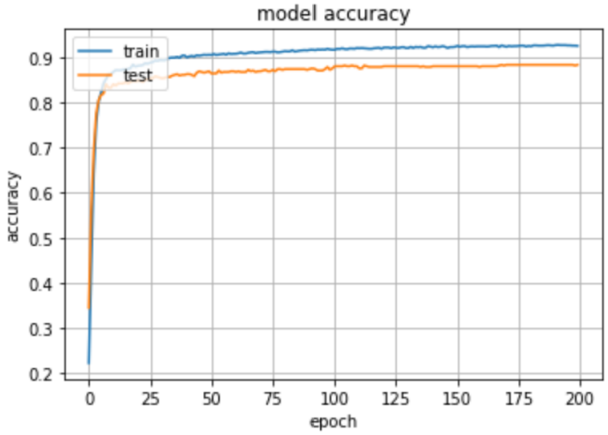

Fig. 6. Évolution du taux de reconnaissance pour un réseau de neurones simple

multicouche, on affine le modèle et on obtient une précision plus élevée avec un taux à 90,4%.

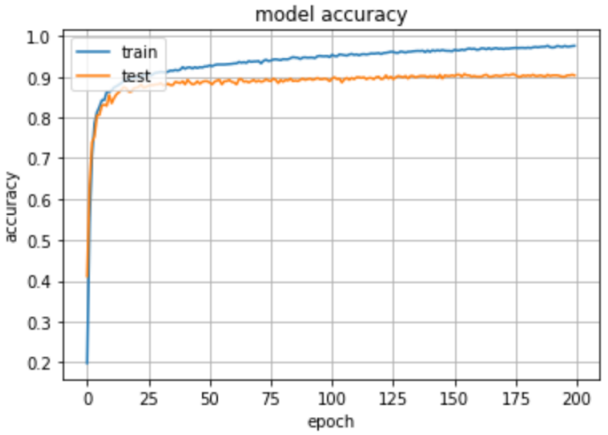

Fig. 7. Évolution du taux de reconnaissance pour un réseau de neurones avec couche supplémentaire de 128 neurones

B. Réseau de neurones CNN

Toutefois, une amélioration est possible en ajoutant un ou plusieurs filtres à convolutions ainsi que des dropout (désactivant des neurones entre epochs et évitant ainsi le sur-apprentissage).

Cette fois ci, 2 filtres à convolutions et un dropout sont rajoutés en amont du MLP. Un dropout de 0.5 est également ajouté entre le MLP et le réseau. En effectuant ces améliorations, on aboutit à la création d'un réseau de neurones convolutionnels (CNN)

Ainsi, le CNN est configuré avec 500 batchs et 256 epochs avec un learning rate de 0.1. Nous obtenons alors la figure 8.

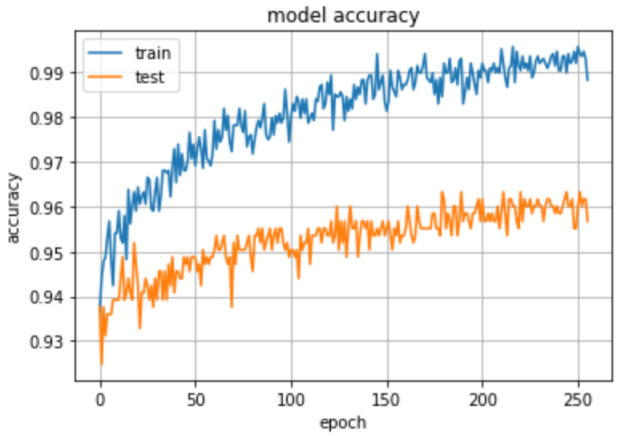

Fig. 8. Évolution du taux de reconnaissance pour un réseau de neurones CNN

La précision est alors à 95.6% pour la base de test, ce qui constitue le taux de reconnaissance le plus élevé de toutes les méthodes explorées précédemment.

VI. Conclusion

Nous avons pu faire un comparatif entre plusieurs méthodes discriminantes pour un problème de classification concret.

Un premier constat est le taux de reconnaissance assez élevé quelque soit la technique de classification choisie, avec pour "pire" résultat le KPPV avec ACP et le réseau de neurones simple avec des taux de reconnaissance resepectifs de 90% et 88%.

Conformement au cahier des charges donné, la solution choisie est donc un réseau convolutionnel de neurones avec un taut de reconnaissance record de pls de 95 %.

Toutefois d'autres indicateurs de performances peuvent être pris en compte pour une application comme la reconnaissance de caractères manuscrits. Une telle application est intéressante en temps réel donc avec un temps d'apprentissage faible et un taux de reconnaissance élevée. Le compromis se trouve alors avec le SVM Multiclasses avec un noyau RBF (2ème meilleur taux de reconnaissance et coût calcul faible).