POLYTECH
PARIS-SUD

Big Data - ET5 - INFO

Rapport de Projet de reconnaissance d'images

Yasmine KERTOUS
Thibault LASOU

Table des matières

Introduction

Le but de ce projet est de concevoir et d'évaluer un système de reconnaissance automatique d'images utilisant différentes méthodes de classification en Python.

Les données du projet sont réparties en 3 catégories : les données d'apprentissage sur lesquelles le système se base pour faire son apprentissage, les données de développement qui serviront à évaluer le système et enfin les données de test. Pour procéder à la classification, on va utiliser : la classification à distance minimum, l'analyse en composantes principales et les support vector machine.
Une fois qu'on a évalué le système en comparant les méthodes de classification, on sélectionne le meilleur système, on lance dessus les données de test et on met dans un fichier le résultat du système.

Dans ce rapport, on va d'abord expliquer nos choix d'implémentation. Ensuite on va comparer et analyser les performances de chacun des classifieurs.

I. Présentation des classifieurs

1. Classifieur à distance minimum (DMIN)

Tout d'abord, le système fait de l'apprentissage. On construit les classes d'images. Chaque classe est représentée par la moyenne des vecteurs d'apprentissage de cette classe. Ensuite pour chacune des classes on calcule un barycentre. Le barycentre de chaque classe va permettre de classifier les images qui font partie des données de développement : on calcule la distance entre le point qui représente l'image et les barycentres. L'image va être classifiée dans la classe dont le barycentre est le plus proche du point qui la représente.

2. Analyse en composantes principales (ACP)

L'analyse en composantes principales est une technique qui permet de réduire la dimension des vecteurs d'images de 784 points à un vecteur de paramètres de plus petite taille.
Pour implémenter cette méthode on a utilisé la librairie Scikit-Learn, qui réduit donc la dimension des vecteurs en fonction de ce qu'on lui demande. Ensuite on classifie les données en calculant les distances entre les barycentres des classes et ces "nouvelles" représentations des images.

3. Support vector machines (SVM)

Les machines à vecteurs de support sont un ensemble de techniques d'apprentissage destinées à résoudre des problèmes de discrimination et de régression. L'algorithme de SVM a pour objectif de trouver la séparation entre les classes et plus la séparation est large plus la séparation est robuste.
Pour implémenter cette méthode, on peut utiliser la même méthode que dans PCA c'est à dire réduire le nombre de composantes. Cela permet de réduire le temps de résolution, en revanche le temps d'apprentissage sera plus long à cause des calculs que suscite la réduction des données.

II. Analyse des résultats

1. Présentation des performances

Pour l'apprentissage on a un échantillon de 10000 données, et 5000 données pour le développement.

Méthode DMIN

Taux d'erreur	Temps d'apprentissage (s)	Temps de résolution (s)
0.3242	0.03127400000000069	0.5094029999999998

Méthode PCA

Nombre de composantes	Taux d'erreurs	Temps d'apprentissage (s)	Temps de résolution (s)
10	0.3462	2.8655	1.8767
20	0.3312	3.7508	1.86404
50	0.3254	7.14667	2.1520
100	0.3246	7.77408	1.26945
180	0.3244	10.38171	1.38344

Méthode SVM

Nombre de composantes	Taux d'erreurs	Temps d'apprentissage (s)	Temps de résolution (s)
sans PCA	0.2204	738.813	0.0655
10	0.2724	4.3904	0.0717
20	0.212	9.3217	0.0752
50	0.175	30.4039	0.1072
100	0.1642	65.4303	0.1028
450	0.1804	1745.8452	0.2251

2. Comparaison et explication des résultats

Comparaison temps d'apprentissage :
La méthode DMIN est plus rapide que PCA car cette méthode prend beaucoup de temps pour calculer la réduction de composantes.
Par ailleurs, pour SVM plus y a de composantes plus c'est lent : comme la réduction de données consomme du temps cela semble naturel que plus y a de composantes plus c'est il y a de calcul donc plus c'est lent.

Comparaison temps d'exécution :
La méthode PCA est plus rapide que DMIN car pour calculer les distances entre les représentation des points et des barycentres, on fait des soustraction, des mises au carrés et une racine carrée. Pour DMIN on fait ces opération sur des données de grande taille 784, en revanche pour PCA, comme on a réduit les tailles des vecteurs, le système effectue donc beaucoup moins de calculs ce qui explique pourquoi PCA est plus rapide.
Néanmoins, SVM est considérablement plus rapide que les 2 autres méthodes.

Pour les taux d'erreur : DMIN a un taux d'erreur inférieur à celui de PCA, et c'est logique parce que dans PCA on réduit la dimension des vecteurs donc on perd de l'information qui permettrait les données entrantes.

Conclusion

DMIN est une méthode de classification supervisée basique permettant de résoudre les problèmes les plus simples.

PCA est une méthode bien adaptée pour traiter de grandes bases de données. En revanche, on perd de l'information en réduisant la dimension des vecteurs ce qui limite sa précision.

La durée d’apprentissage de SVM est plus longue par rapport aux deux autres. Par ailleurs, ses performances sont nettement meilleures que les 2 autres méthodes.