International Journal of Scientific Research & Engineering Technology (IJSET)
Vol.7 pp.1-5

CLASSIFICATION DES ESPECES VEGETALES PAR FAMILLE

Trey Zacrada Françoise Odile[#1], Konan Brou Marcellin[#2], E.Olajubu[*3]
[#]*Electronique et Electricité Appliquées, Institut National Polytechnique Houphouet Boigny Yamoussoukro, Cote d'Ivoire*
[*] Faculty of Computer Science Obafemi Awolowo University
Ile-Ife, Nigeria
[1]mariefranceodiletrey@gmail.com
[2]konanbroumarcellin@yahoo.fr
[3]emmolajubu@oauife.edu.ng

***Abstract*— La classification a été au cœur de plusieurs débat chez les botanistes. Cela vient de la difficulté à pouvoir définir clairement les différentes familles des plantes. En effet, la majorité des méthodes utilisées sont manuelles ce qui rend cette tache assez difficile. A travers cet article nous faisons recours à la classification automatique afin de faciliter le regroupement des plantes par similarité. Pour ce faire, le modèle ZACclassification, contenant l'algorithme de K-Means est utilisé dans l'optique de classifier les plantes en famille.**

***Keywords*— Apprentissage automatique, taxonomie, Vegetable dataset, K-means.**

I- INTRODUCTION

La préservation de la biodiversité repose sur une classification précise, fondée sur la science (autrement dit, un système de désignation des organismes). En effet, sans cela, nous serons incapables de décrire la multitude d'espèces qui peuplent les forêts tropicales, et de les comparer au petit nombre qui vivent dans nos pays. En l'absence d'une telle classification, il serait impossible d'identifier les espèces végétales de notre environnement [1].

A- Etat de l'art

Traditionnellement, les botanistes utilisent les clés d'identification pour classer les plantes. Ces identifications concernent généralement les caractéristiques morphologiques des végétaux. [2] [3] [4]. Celle-ci sont exploitées manuellement et concerne principalement les caractéristiques telles que : la feuille, la tige, la fleur, les fruits et les racines. Leur mode opératoire reste une activité difficile. Pour palier à cette insuffisance, certains acteurs procèdent à la digitalisation de ces clés.

Ils ont créé un outil de conversion des documents papiers en dossiers électroniques. Mais cet outil n'a pas été validé par des experts [5], par conséquent, il est peu fiable. C'est sur cette lancée, qu'un outil polyvalent de base de données informatisée des plantes se constitue [6].

Par ailleurs *Thierry Pernot et al.* se servent de l'apprentissage automatique pour identifier une plante à partir de sa feuille. Ils utilisent la photographie de cette feuille pour extraire les caractéristiques telles que la texture, la couleur et la forme tout en s'appuyant sur l'algorithme de réseaux de neurones convolutionels [7]. Cependant, ces descripteurs ne donnent pas de sémantique particulière sur l'image [8]. Aussi, force est de constater que dans son processus d'identification, la segmentation qui devrait augmenter la précision de la classification, biaisait les résultats [9].

B- Méthodologie

À travers cet article, nous proposons un mode de classification plus poussée en utilisant les descripteurs de la hauteur de la feuille ainsi que celui de la taille de la tige. L'accent sera mis sur leurs valeurs réelles et non celles extraites de leur image. Pour ce faire, nous nous servons de l'algorithme de *K-Means* pour la mise en place de cette classification non supervisée en vue de regrouper nos plantes en clusters de famille.

ISSN 1737-9296

Ce modèle nommé ZACClassification se déroule en trois étapes : l'élaboration de la base de données, le pré-traitement et la classification. Cette approche est illustrée par la *figure 1* suivante :

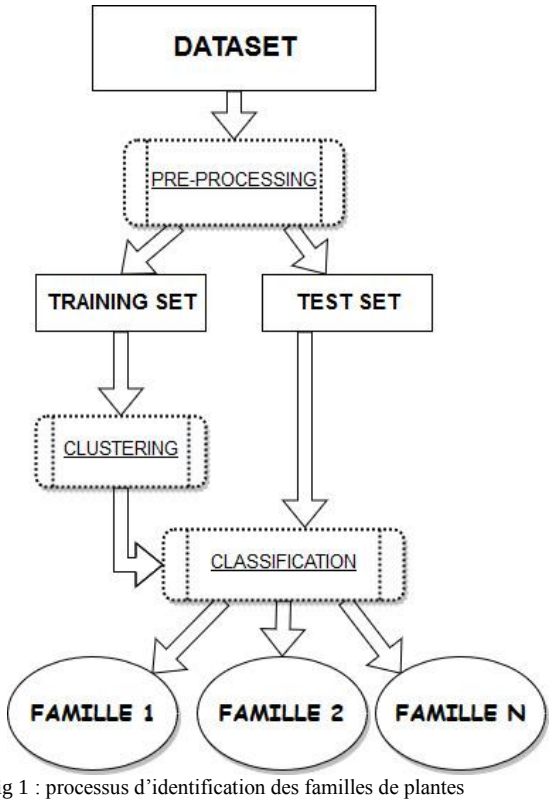

Fig 1 : processus d'identification des familles de plantes

Pour mettre en pratique cette méthodologie, nous modélisons toutes les données afférentes à nos plantes, ensuite, nous simulons les résultats les interprétons et enfin une conclusion.

II- MODELISATION

A- Elaboration du dataset

Pour réaliser notre dataset, nous faisons l'acquisition des données de différents sources et procédons à son épuration. La *figure 2* ci-dessous nous présente le mode de fonctionnement.
Selon nos sources d'acquisition des différentes plantes [10] [11] [12], nous constituons une base de données. Nous reportons le nom de la plante aves la taille maximale de sa tige (TMAX) en millimètres et la longueur maximale de sa feuille (LMAX) en millimètre. C'est la base de données redondantes comme l'indique la figure suivante

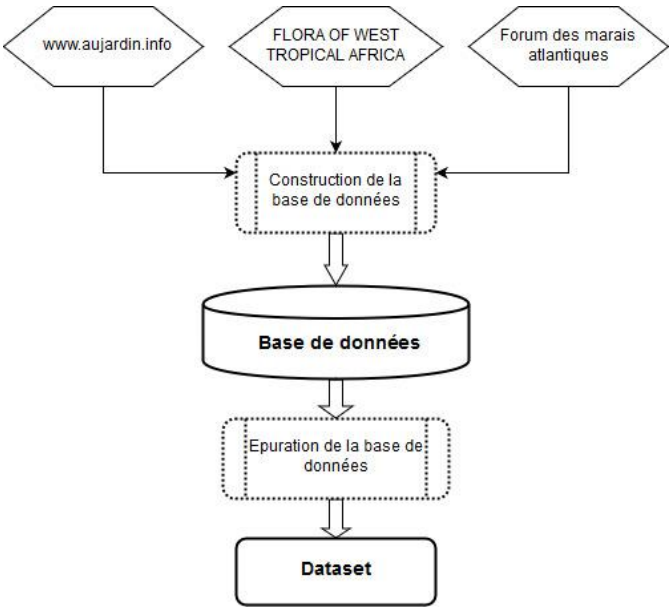

Fig 2. Elaboration du Dataset

Tableau 1: Base de données redondantes (J. Hutchinson et al)

PLANTES	TMAX	TMIN	LMAX	LMIN
Thomsonii	13716	13716	150	100
Xylopia Africana	12192	9140	160	90
Staudtii	45720	45720	160	90
Ruberscens	27432	27432	240	90
Eliotii	9144	27432	90	50
…	…	…	…	…
Sofa	6096		90	60
Afzelii	4572		250	70
Margaritaceus	1200		4	4
Aucheri	1500		5	4
Uvaria Scabrida	9144		180	100

Après épuration de cette base, nous obtenons le dataset suivant :

ISSN 1737-9296

Index	TMIN	TMAX	LMIN	LMAX
0	1.37e+04	1.37e+04	100	150
1	9.14e+03	1.22e+04	90	160
2	4.57e+04	4.57e+04	90	160
3	2.74e+04	2.74e+04	90	240
4	2.74e+04	2.74e+04	130	130
5	9.14e+03	9.14e+03	50	90
6	2.44e+04	2.44e+04	60	130
7	1.83e+04	1.83e+04	150	150
8	914	2.44e+03	40	80
9	3.05e+03	3.05e+03	45	150
10	6.1e+03	6.1e+03	70	120
11	6.1e+03	6.1e+03	60	90
12	4.57e+03	4.57e+03	70	250
13	9.14e+03	9.14e+03	100	180

Fig 1: Dataset PlantData

B- Pré-traitement

Nous débarrassons notre base de données de toutes les données manquantes et catégorielles. Nous mettons toutes ces données à la même échelle. Le traitement ainsi terminé, nous divisons le data set en training set et test set

	0	1
0	1500.00	3.50
1	20.00	10.00
2	36576.00	240.00
3	1000.00	2.00
4	1500.00	0.90
5	9144.00	180.00
6	10.00	2.50
7	1000.00	2.00
8	160.00	7.00
9	9144.00	250.00

Fig 2: Training Set

	0	1
0	2438.40	80.00
1	45.00	1.50
2	1200.00	8.00
3	1300.00	2.00
4	450.00	90.00
5	9144.00	500.00
6	130.00	5.00
7	27432.00	160.00
8	170.00	40.00
9	12192.00	180.00

Fig 3: Test Set

C- Classification

De tous les algorithmes du clustering, le *K-MEANS* est le plus utilisé. Il a la capacité de bien analyser un ensemble de données, caractérisés par des descripteurs afin de regrouper ces données en clusters [13].

Étant donné un ensemble de plantes $(P_1, P_2, \ldots, P_n)$. On cherche à partitionner les n plantes en K familles. $\{F=F_1, F_2 \ldots, F_n\}(K<n)$. En minimisant la distance (d) entre les plantes à l'intérieur de chaque partition (ou distance inter-cluster) $\text{Arg min}\sum\sum \|P_j - \mu_i\| = \text{Arg min}\sum \|F_i\|$ var F_i, ou μ_i, la moyenne des plantes de famille.

III- SIMULATION

Nous considérons notre *dataset, PlantData* composé de 129 plantes. Nous utilisons le training set composé de 103 plantes pour l'entrainement et réservons les 26 plantes du test set pour tester notre modèle.

A- Déterminations des K familles ou clusters

Nous utilisons la méthode *Elbow* [14], pour trouver le nombre optimal de clusters.

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
 kmeans = KMeans(n_clusters = i,
init = 'k-means++', random_state = 0
 kmeans.fit(X_test)
wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('La méthode Elbow')
plt.xlabel('Nombre de clusters')
plt.ylabel('WCSS')
lt.show()
```


ISSN 1737-9296

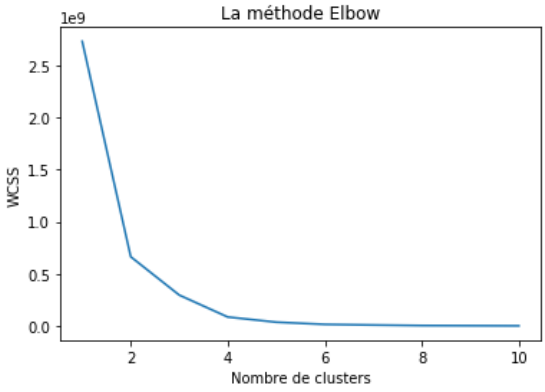

Fig 4: Choix du cluster par la méthode Elbow

Visualisons nos plantes dans le plan avant l'application du l'algorithme du K-Means.

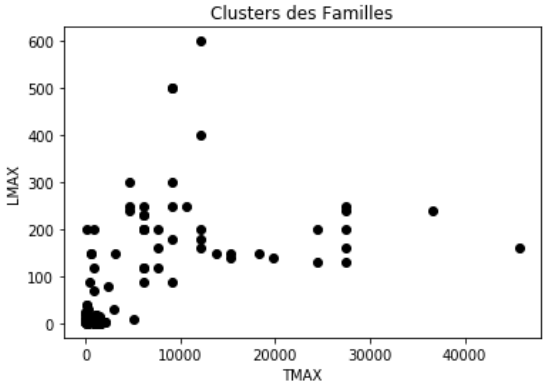

Fig 5: Les différentes plantes à classifier

- Algorithme de la construction du modèle

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 3, init =
'k-means++', random_state = 0)
y_kmeans = kmeans.fit_predict(X_test)
plt.scatter(X_test[y_kmeans == 1, 0],
X_test[y_kmeans == 1, 1], c = 'red',
label = 'Cluster 1')
plt.scatter(X_test[y_kmeans == 2, 0],
X_test[y_kmeans == 2, 1], c = 'blue',
label = 'Cluster 2')
plt.scatter(X_test[y_kmeans  ==  0,  0],
X_train[y_kmeans == 0, 1],  c = 'green',
label = 'Cluster 3')
plt.title('Clusters des Familles')
plt.xlabel('TMAX')
plt.ylabel('LMAX')
plt.legend()
```

Appliquons le modèle à nos plantes à classifier

Nous obtenons des clusters de familles comme l'indique la figure suivante :

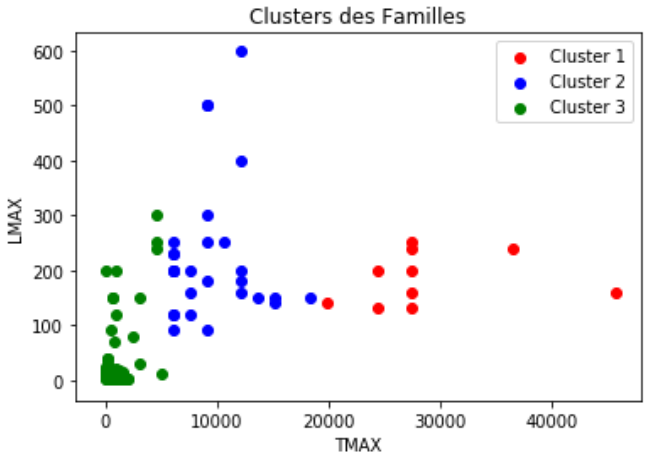

Fig 6: Famille des plantes

B- Performance du modèle

A partir des plantes du test set, nous évaluons la performance de notre modèle. Le tableau ci-dessous récence les plantes prédites à partir de celles qui ont été observées.

Tableau 2 : la matrice de confusion (Philippe Beraud)

	PLANTES PREDITES			
PLANTES OBSERVEES	**CLUSTERS**	**1**	**2**	**3**
	1	6	0	2
	2	0	8	2
	3	2	0	7

Dans le tableau suivant, nous calculons les indicateurs de base de la qualité de la prédiction sur les différents clusters.

Tableau 3 : les indicateurs de performance (Philippe Beraud)

	PRECISION	**RAPPEL**	**F-mesure**
CLUSTER 1	0.857142	0.75	0.8
CLUSTER 2	0.8	1	0.88889
CLUSTER 3	0.777778	0.7	0.736842

La précision totale du modèle s'élève à 81,16% avec une marge d'erreur de 18,83%

C- Interprétation

Selon le critère de la longueur maximale de la feuille et la taille maximale de la tige des plantes, nous obtenons trois familles différentes de plantes, à savoir *le cluster 1* représenté par la

ISSN 1737-9296

couleur rouge, *le cluster 2*, la couleur verte et enfin, *le cluster 3* par la couleur bleue.
Les résultats ont été simulés avec un training set composé de 103 plantes. Nous avons testé le modèle avec une base de 26 plantes et les résultats obtenus sont conformes aux résultats de l'entrainement. De plus, par analogie à la réalité, *le cluster 1* regroupe les arbres dont la longueur des feuilles varie entre 150 et 250 millimètres. Il correspond à la famille des Annonacea.

CONCLUSION

Au terme de cette étude, nous avons défini la taille de la tige et la longueur des feuilles des plantes comme descripteurs pertinents de famille de plantes. Aussi avons-nous développé un modèle de classification, basé sur l'algorithme de *K-Means*. Pour une base de 129 plantes, ce modèle a pu regrouper de façon automatique ces espèces végétales en trois familles distinctes. Le taux de réussite de cette classification est de 81,16%

REMERCIEMENT

The World Bank ICT Department, Faculty of Computer Science and Faculty of Botany Obafemi Awolowo University. Le LARIT pour les moyens mis à disposition pour le travail effectué.

Nous voulons aussi remercier en particulier Mr Aderounmu G. S., Mr Goore Bi Tra, Mrs Odukoya, Mr N'Guessan Behou.

REFERENCES

[1] Dominique Picouet et Al, les règles de la taxonimie : nommer les espèces 2018

[2] Denis Paquette, clés des 16 genres de Cypéraceaes, Flora Quebeca, 19 Septembre 2016

[3] André Sabourin, clé des crucifères., flora québéca, mars 2018

[4] Denis Paquette, clés desvergesd'or, Flora Québéca, 2016

[5] Rousse Guillaume, Éric Villemonte de La Clergerie. Analyse automatique de documents botaniques: le projet Biotim. proc. of TIA'05 : Journées Terminologie ; Intelligence Artificielle, Apr 2005, Rouen, France, France. 2005

[6] Raymond Boyd et Al, Une base de données informatisée transdisciplinaire de la flore chez les sémé du burkina faso : un outil pour l'étude du lien nature-société, 2014

[7] Thierry Piernot et Al, Flora Bellissima, un nouvel outil pour découvrir la flore, mars 2014.

[8] Zhong-Qiu Zhao et al), Apleaf : an efficient android-based plant leaf identification system, Neurocomputing, 2014

[9] Mads Dyrmann et al., 2016, Plants species classification using deep convolutional neural network, 10.1016/j.biosystèmseng.2016.08.024, 2016

[10] *Forum des Marais Atlantiques 2017,Herbier Numérique, Flore en zonz humide*, Région nouvelles d'Aquitaine, 93 pages http:// www.forum-zones-humides.org

[11] *Schoonderwoerd et al, Zygotic dormancy underlies prolonged seed development in Franklinia alatamaha (Theaceae): a most unusual case of reproductive phenology in angiosperms, Botanical Journal of the Linnean Society, Volume 181, Issue 1, 1 May 2016, Pages 70–83.*

[12] *Hutchinson, J., Dalziel, J.M., Keay, R.W.J., Hepper, N.: Flora of West Tropical Africa. (2014).*

[13] *M. Dundar, Q. Kou, B. Zhang, Y. He, and B. Rajwa, "Simplicity of Kmeans Versus Deepness of Deep Learning: A Case of Unsupervised Feature Learning with Limited Data," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 883–888.*

[14] *Bholowalia, P., Kumar, A.: EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications. 105, (2014).*

[15] *Philippe Beraud - MSFTAugust 5, (2014)*

ISSN 1737-9296