International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Etude Portant Sur La Combinaison De Classifieurs Pour Augmenter Le Degré D'homogénéité Des Individus D'un Ensemble

Pierrot Mukendi Ngalamulume

Ct, Ises Kananga

Résumé

Face à une grande masse des données, ces dernières doivent être analysées pour en faire sortir des informations fiables à la prise de décision. Elles peuvent être regroupées en sous-ensembles selon les critères bien définis. Ainsi, La force de la classification réside dans sa capacité à répartir les objets en classes distinctes dont les éléments ont les mêmes propriétés. Or, il n'existe pas le meilleur classifieur capable de traiter n'importe quelle distribution des données d'apprentissage, ni discriminer correctement un ensemble important de classes. D'où, il s'avère toujours important de les combiner pour la précision et l'efficacité des résultats. Ceci offre plusieurs avantages tels que la distribution des caractéristiques sur des classifieurs adaptés, l'exploitation de la complémentarité entre les classifieurs, la prise en compte des performances de chacun des classifieurs.

C'est dans ce cadre que s'inscrit cette étude qui porte sur la combinaison séquentielle des classifieurs pour augmenter le degré d'homogénéité des individus d'un ensemble. Deux classifieurs seront combinés, l'arbre de décision de l'apprentissage supervisé et le K-means de l'apprentissage non supervisé pour constituer des groupes beaucoup plus homogènes d'individus.

Abstract

Faced with a large mass of data, it must be analyzed to provide reliable information for decision-making. They can be grouped into subsets according to well-defined criteria. Thus, the strength of classification lies in its ability to divide objects into distinct classes whose elements have the same properties. However, there is no best classifier that can handle any distribution of training data, nor can it correctly discriminate a large set of classes. Hence, it is always important to combine them for the accuracy and efficiency of the results. This offers several advantages such as the distribution of characteristics on adapted classifiers, the exploitation of the complementarity between the classifiers, the taking into account of the performance of each of the classifiers.

It is in this context that this study focuses on the sequential combination of classifiers to increase the degree of homogeneity of the individuals in a set. Two classifiers will be combined, the decision tree of supervised learning and the K-means of unsupervised learning to form much more homogeneous groups of individuals.

1. Introduction

A l'ère de la numérisation et de la digitalisation, le monde des données connaît une véritable révolution s'adaptant aux nouveaux besoins des entreprises. Pour prendre de meilleures décisions au sein de ces

dernières, les données et, en particulier leur analyse, offrent un avantage stratégique.
Elle est un atout précieux pour les entreprises lors d'une prise de décision importante. Pour y parvenir, plusieurs techniques d'analyse sont utilisées selon qu'il s'agit de l'apprentissage supervisé, non supervisé ou incrémental. Toutes ces techniques fournissent des résultats. C'est le cas de la classification, regroupement, association, modèle de suivi, réseau de neurones, arbre de décision, etc.
Cependant, dans le cas de la segmentation de données, nous constatons qu'il n'existe pas le meilleur classifieur capable de traiter n'importe quelle distribution des données d'apprentissage, ni discriminer correctement un ensemble important de classes. D'où, il s'avère toujours important de les combiner pour la précision et l'efficacité des résultats. Ceci offre plusieurs avantages tels que la distribution des caractéristiques sur des classifieurs adaptés, l'exploitation de la complémentarité entre les classifieurs, la prise en compte des performances de chacun des classifieurs, en rejoignant le principe de diviser pour mieux régner.
C'est dans ce cadre que s'inscrit cette étude qui porte sur la combinaison séquentielle des classifieurs pour augmenter le degré d'homogénéité des individus d'un ensemble. Deux classifieurs seront combinés, l'arbre de décision de l'apprentissage supervisé et le K-means de l'apprentissage non supervisé pour constituer des groupes beaucoup plus homogènes d'individus.
Dans la suite de cette étude, nous aurons à définir le classifieur, présenter ses différents types de sortie et ses mesures de performance, ensuite nous allons montrer comment construire un système multi classifieur, puis poser notre problématique et enfin élaborer notre architecture du système en utilisant une méthode de combinaison séquentielle, présenter notre algorithme de deux classifieurs combinés et une simulation par application aux données médicales.

2. Définitions

a. Classifieur

Un classifieur est un outil de reconnaissance qui reçoit une entrée (forme, document, mot, etc.), et fournit en sortie des informations concernant cette forme inconnue. C'est à dire, pour une entrée donnée (x), le classifieur (e) lui attribue une classe (Ci, tq i=1..m), parmi les (m) classes existantes.[1]

b. Types de sorties du classifieur[2]

Suivant le niveau d'information apporté par le classifieur, on peut distinguer quatre catégories de sorties (Laurent, 1999) :

- Sortie de type classe : C'est le type le plus général mais qui apporte le moins d'informations. Le classifieur ne donne que la proposition du type (la classe) de l'entrée à reconnaître sans aucune autre information ej(x) = Ci (i ϵ {1..m}).
- Sortie de type ensemble Le classifieur donne sa réponse sous forme d'un ensemble de classes candidates sans préciser ses préférences ej(x) ={Ci / i <= m}.
- Sortie de type rang Ce type de sortie reflète l'ordre de préférence des propositions fournies par le classifieur, cela se traduit par l'attribution d'un rang pour chaque classe ; Plus cette dernière est probable moins le rang est élevé. La liste des propositions peut contenir toutes les classes possibles ou

[1] Gasmi. I et alii, *Combinaison de classifieurs*, 3rd International Conference : Sciences of Electronic, Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA
[2] Laurent M., ''Combinaison de classifieurs pour la reconnaissance de formulaires structurés''. DESS/DEA, Université de Rouen (1999).

seulement les mieux classées. ej(x) =[rj 1, rj 2, …, rj m] où rj i est le rang attribué à la classe (i) par le classifieur (j).

- Sortie de type mesure Cette sortie est la plus riche en informations puisque le classifieur dans ce cas associe à chaque classe une mesure de confiance qui peut être, par exemple, une probabilité. ej(x)=[Mj 1, Mj 2,….,Mj m] où Mj i est la mesure attribuée à la classe (i) par le classifieur (j).

c. Mesures de performances d'un classifieur[3]

Pour une entrée donnée, un classifieur peut générer les réponses suivantes :

- Un rejet : pour indiquer que le classifieur n'a pas pu identifier cette entrée.
- Une reconnaissance : dans ce cas, il identifie bien l'entrée, et il lui attribue sa classe appropriée.
- Une substitution : le classifieur attribue une autre classe à l'entrée.

La performance d'un classifieur peut être mesurée en calculant les trois taux suivants :

Taux de rejet=$\frac{Nombre\ de\ formes\ rejetées}{Nombre\ total\ de\ formes}$

Taux de reconnaissance=$\frac{Nombre\ de\ formes\ reconnues}{Nombre\ total\ de\ formes}$

Taux de substitution=$\frac{Nombre\ de\ formes\ males\ reconnues}{Nombre\ total\ de\ formes}$

d. Construction d'un système multi-classifieurs [4]

Un système multi-classifieurs (Multiple Classifier System : MCS) est constitué d'un ensemble de différents classifieurs et d'une fonction de décision pour combiner leurs sorties. La description d'un MCS suit les deux phases suivantes :

- Générer un ensemble de classifieurs complémentaires qui peuvent être combinés pour arriver à une solution optimale.
- Définir la fonction de combinaison pour donner une décision finale.

La difficulté de choix des classifieurs a poussé les chercheurs à développer des méthodes pour aider les concepteurs à effectuer ce choix. Parmi ces méthodes, celle de « test et sélection » (over produce and choose paradigm). L'idée principale de cette méthode est de produire un ensemble initial large de classifieurs candidats, puis sélectionner un sous ensemble qui est jugé le plus valable pour aboutir à des performances optimales. Pour le faire, le procédé suit deux cycles :

- Construire l'ensemble de classifieurs de départ (over production).
- Choisir le sous ensemble le plus intéressant.

Le choix de la fonction de décision joue un rôle très important dans la conception d'un MCS. La fonction de décision peut être conçue comme étant une fonction de combinaison, par conséquent la sortie du MCS reflète la décision de tout l'ensemble en utilisant par exemple le vote majoritaire, la somme pondérée, etc. Ou bien comme étant une fonction de sélection dynamique d'un classifieur, dans ce cas il faut avoir au moins un classifieur dans l'ensemble qui pourra classer correctement une forme d'entrée.

Le mécanisme de sélection sera plus efficace lorsque les classifieurs sont 'spécialisés', d'où la facilité de calcul des conditions de sélection affectant chaque forme d'entrée au classifieur le plus convenable. Par contre, le mécanisme de combinaison sera plus efficace lorsque les classifieurs manifestent des comportements différents.

[3] Gasmi. I et alii, op cit.

[4] Azizi & al., *''Un système multi-classifeurs neuronaux à combinaison floue pour la reconnaissance de l'écriture arabe manuscrite''*. MCSEAI'2002, septième conférence magrébine en Informatique, Annaba, Algérie (2002).

3. Problématique de la combinaison séquentielle des classifieurs

Le problème de la combinaison séquentielle de classifieurs peut se poser de la façon suivante : Disposant d'un ensemble de N classifieurs, comment les combiner pour homogénéiser les individus et fiabiliser une prise de décision ?

Tout cela dépendra de la façon dont on veut faire interagir les classifieurs, soit Indépendamment les uns des autres (vote ?), soit par Elimination d'hypothèses (décisions dépendantes) ou soit par Coopération de classifieurs (chacun résout un problème…).

Quant à nous, nous allons les faire interagir les uns indépendamment des autres.

4. Architecture du système

L'objectif principal de notre travail est de concevoir et réaliser une combinaison de classifieurs segmentaux pour augmenter le degré d'affinités entre individus d'un même ensemble de données.

L'idée principale est d'utiliser deux algorithmes de segmentation dont le premier est l'arbre de décision qui permettra de créer des groupes homogènes partant d'un ensemble de données et le deuxième K-means qui sera appliqué sur les groupes déjà triés (résultats de l'arbre de décision) pour constituer de groupes plus homogènes partant de nouveaux critères établis.

L'étape de combinaison consiste à prendre une décision finale en appliquant des fonctions de combinaison sur les résultats de classification donnés par la technique de l'arbre de décision et la technique K-means.

5. Les méthodes de combinaison des classifieurs[5]

L'intérêt croissant des chercheurs pour la combinaison des classifieurs, a entraîné la mise au point de nombreux schémas traitant les données de manières différentes. Généralement, trois approches pour la combinaison de classifieurs peuvent être envisagées : parallèle, séquentielle et hybride. Cependant, malgré l'existence d'un grand nombre de des schémas de combinaison, la détermination de la meilleure architecture reste un problème ouvert. L'approche que nous avons adopté pour notre système est l'approche séquentielle.

a. Approche séquentielle ou série

La combinaison séquentielle, appelée également combinaison série, est une architecture organisée en niveaux successifs de décision qui servent à réduire d'une manière progressive le nombre de classes possibles (La liste de classes candidats est ainsi progressivement réduite jusqu'à ce qu'il ne reste qu'une seule décision possible), où chaque niveau de décision est caractérisé par la présence d'un seul classifieur qui doit prendre en compte la réponse fournie par le classifieur placé en amont, pour traiter un rejet ou confirmer une décision obtenue sur la forme présentée à son entrée.

Cette approche peut être considérée comme un filtrage progressif des décisions, dans la mesure où elle permet de diminuer au fur et à mesure l'ambiguïté sur la classe proposée. Cela permet généralement de diminuer le taux d'erreur globale de la chaîne de reconnaissance. Néanmoins, l'inconvénient majeur de cette approche de combinaison, réside dans sa sensibilité à l'ordre d'exécution des classifieurs, qui est primordial et influe sur le résultat final.

b. Combinaison séquentielle de classifieurs.

En effet, les classifieurs placés en amant doivent être robustes, c'est-à-dire que la vraie classe de la forme à reconnaitre doit apparaître dans les listes successives proposées. Car un échec de décision du premier

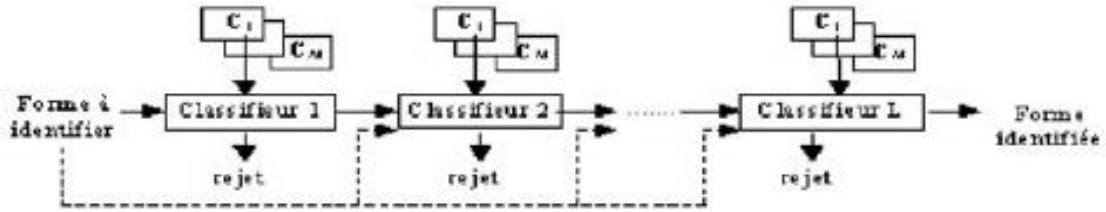
[5] Les méthodes de combinaison des classifieurs (123dok.net), dans le document Les technique de reconnaissance de formes application a la reconnaissance de l'écriture arabe (Page 93-96), consulté le 14/01/2024 à 23h09'

classifieur provoque une erreur, qui va se propager de façon irréversible dans toute la série des classifieurs utilisés. Donc la combinaison séquentielle, exige une certaine connaissance a priori du comportement de chacun des classifieurs. Dans cette approche, chaque classifieur est ajusté en fonction du classifieur placé en amont, ce qui peut conduire à un ré-paramétrage des classifieurs suivants en cas d'une simple modification du premier classifieur.

Dans cette approche, on trouve l'organisation en niveaux successifs de décision permettant de réduire progressivement le nombre de classes possibles

A chaque niveau : un seul classifieur prend en compte la réponse fournie par le classifieur placé en amont pour :

- Traiter les rejets
- Confirmer la décision obtenue à l'étage précédent



+ Filtrage progressif des décisions (réduction de l'ambiguité)
- Sensible à l'ordre dans lequel sont placés les classifieurs
- Suppose une connaissance a priori du comportement de chacun des classifieurs
- Difficile d'optimiser l'ensemble (application-dependent)

6. Combinaison de l'arbre de décision et K-means

a. Principes de fonctionnement Arbre de décision & K-means

1. Déterminer la meilleure caractéristique dans l'ensemble de données
2. Diviser les données en sous-ensembles contenant les valeurs possibles de la meilleure caractéristique
3. Générer de manière récursive de nouveaux sous arbres de décision en utilisant les sous arbres de données créés
4. S'arrêter lorsqu'on ne peut plus classifier
5. ***Sortie :*** *on obtient de différentes classes de données. On choisit une classe parmi celles obtenues en sortie et on applique le K-means*
6. Estimer aléatoirement des points K parmi les données de cette classe
7. Assigner les éléments à ces groupes K
8. Déplacer les points K vers les centres
9. Réassigner les éléments et répéter jusqu'à stabilité

b. Algorithme de la combinaison Arbre_K-means

Entrée :

Ensemble de N données, noté par x

Nombre de groupes souhaité, noté par k

Sortie :

Une partition de K groupes {C1, C2,…, Ck}

Début

ArbreDecision (T)

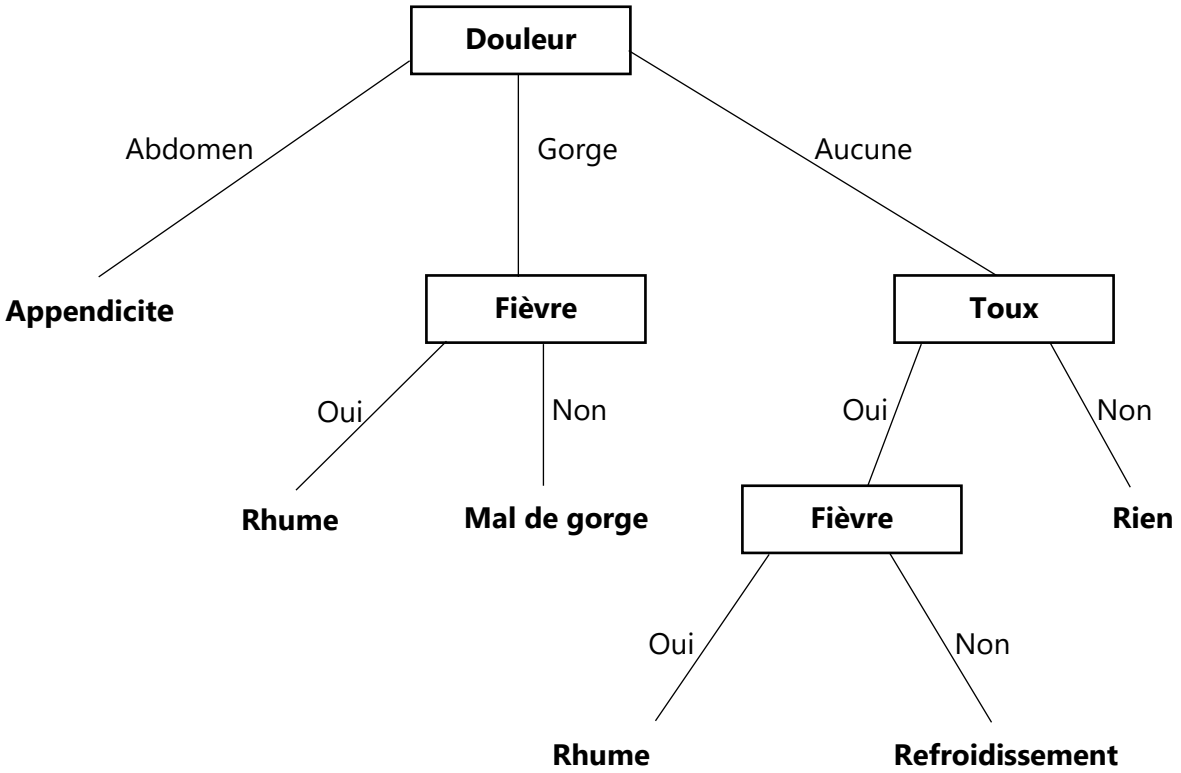
Si condition d'arrêt alors
Retourner feuille (T)
Sinon
Choisir le meilleur attribut 1 entre 1 et m
Pour chaque valeur v de l'attribut 1
T[v]={(x,y) de T tels que x-1=v}
t[v]=ArbreDecision(T[v])
Fin Pour
Retourner nœud (1, {v t[v]})
Fin si
Initialisation aléatoire des centres Ck
Répéter
Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche
Représentation : calculer les centres associés à la nouvelle partition
Jusqu'à convergence de l'algorithme vers une partition stable
Fin

7. Application aux données médicales

Soit l'ensemble de données suivant :

N°	FIEVRE	DOULEUR	TOUX	MALADIE
01	Oui	Abdomen	Non	Appendicite
02	Non	Abdomen	Oui	Appendicite
03	Oui	Gorge	Non	Rhume
04	Oui	Gorge	Non	Rhume
05	Non	Gorge	Oui	Mal de gorge
06	Oui	Non	Non	Aucune
07	Oui	Non	Oui	Rhume
08	Non	Non	Oui	Refroidissement
09	Non	Non	Non	Aucune

De cet ensemble découle l'arbre de décision suivant :



Ces résultats sont classés déjà dans 5 groupes (Appendicite, Rhume, Mal de gorge, Refroidissement, Aucune maladie).
Sur base de ces résultats fournis par cet arbre de décision, nous pourrions sélectionner une classe, par exemple le groupe « Appendicite » et appliquer K-means en ajoutant des attributs tels que la tranche d'âge et le genre.

En ajoutant les données de la tranche d'âge, nous aurons le tableau suivant :

N°	FIEVRE	DOULEUR	TOUX	MALADIE	Tranche d'âge
01	Oui	Abdomen	Non	Appendicite	<18
02	Non	Abdomen	Oui	Appendicite	18 à 45
03	Oui	Gorge	Non	Rhume	<18
04	Oui	Gorge	Non	Rhume	>45
05	Non	Gorge	Oui	Mal de gorge	18 à 45
06	Oui	Non	Non	Aucune	>45
07	Oui	Non	Oui	Rhume	<18
08	Non	Non	Oui	Refroidissement	18 à 45
09	Non	Non	Non	Aucune	<18

En appliquant K-means sur l'ensemble d'éléments du groupe « Appendicite », nous aurons 3 classes suivantes avec les individus plus homogènes :

- Classe de la Tranche (<18)
- Classe de la tranche (18 à 45)

- Classe de la tranche (>45)

Si l'on voulait les classer selon le genre partant du tableau ci-dessous :

N°	FIEVRE	DOULEUR	TOUX	MALADIE	Genre
01	Oui	Abdomen	Non	Appendicite	Masculin
02	Non	Abdomen	Oui	Appendicite	Féminin
03	Oui	Gorge	Non	Rhume	Féminin
04	Oui	Gorge	Non	Rhume	Masculin
05	Non	Gorge	Oui	Mal de gorge	Féminin
06	Oui	Non	Non	Aucune	Masculin
07	Oui	Non	Oui	Rhume	Féminin
08	Non	Non	Oui	Refroidissement	Masculin
09	Non	Non	Non	Aucune	Féminin

Nous aurons pour l'ensemble d'éléments du groupe « Appendicite » deux classes :

- Classe des hommes : 1 élément
- Classes de femmes :1 élément

En considérant le groupe des individus de l'Appendicite par tranche d'âge et genre nous aurons les classes les plus homogènes suivantes :

1. **Tranche <18**
   - Masculin :1 élément
   - Féminin : 0 élément
2. **Tranche 18 à 45**
   - Masculin : 1 élément
   - Féminin : 0 élément
3. **Tranche >45**
   - Masculin : 0 élément
   - Féminin : 0 élément

Conclusion

Il était question dans cette étude de combiner les techniques de data mining dans l'exploration de données pour constituer des groupes plus homogènes. Deux classifieurs segmentaux ont été choisis dont l'arbre de décision et le K-means. La combinaison de ces deux est un outil efficace pour augmenter la performance et fournit une grande exactitude de classification.

Quant à la méthode de combinaison, l'approche séquentielle a été de mise, l'arbre de décision est appliqué sur les données issues d'un ensemble X et a produit des résultats. Ici les données sont réparties dans différents groupes. Le décideur voulant approfondir ses analyses, peut encore prendre ces résultats et les soumettre à d'autres traitements. On choisit un groupe de données considéré comme un ensemble et on applique K-means, on obtiendra des groupes avec des données beaucoup plus homogènes.

Cette étude pourrait être complétée par l'utilisation de plus de deux classifieurs avec des méthodes de combinaisons différentes.

Bibliographie

1. (Azizi & al., 2002) Azizi N., Sari T., Souici L., Sellami M., ‘‘*Un système multi-classifeurs neuronaux à combinaison floue pour la reconnaissance de l’écriture arabe manuscrite’’*. MCSEAI’2002, septième conférence magrébine en Informatique, Annaba, Agérie (2002).
2. (Laurent, 1999) Laurent M., *‘‘Combinaison de classifieurs pour la reconnaissance de formulaires structurés’’*. DESS/DEA, Université de Rouen (1999).
3. (Souici & al., 2000) Souici L., Aoun A., Sellami M., *‘‘Vers une architecture multi-classifieurs pour la reconnaissance de montants de chèques arabes’’*. Maghrebian Conference on Software Engineering and Artificial Intelligence, MCSEAI’2000, Fès, Morocco (2000).
4. (Zouari & al., 2002) Zouari H., Heutte L., Lecourtier Y., Alimi A., *‘‘Un panorama des méthodes de combinaison de classifieurs en reconnaissance de formes’’*. RFIA2002, 11th congrès francophone de AFRIF-AFIA de reconnaissance des formes et Intelligence Artificielle, pp : 499-508, Angers (2002)
5. (XU & al., 1992) Xu l., Krzyzak A., Suen C. Y., *‘‘Methods of combining multiple classifieurs and their application to handwriting recognition’’*. IEEE transactions on systems, man and cybernetics, vol 22, N:0 3, pp: 418-435, (1992).
6. (Montoliu, 1995) Montoliu L., *‘‘Architecture multi agents et réseaux connexionnistes : Application à la lecture de chèques manuscrits’’*. Thèse de Doctorat, Laboratoire LIX, Ecole polytechnique, Palaiseau, France (1995).
7. Gasmi. I et alii, *Combinaison de classifieurs*, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA