Classification d'Images de Documents avec Retour de Pertinence : Application aux Documents de Type Ressources Humaines

Olivier AUGEREAU, Nicholas JOURNET, Jean-Philippe DOMENGER

Laboratoire Bordelais de Recherche en Informatique
351 cours de la Libération, 33405 Talence, France
{olivier.augereau,nicholas.journet,jean-philippe.domenger}@labri.fr

Résumé – Cet article aborde le problème de classification d'images de documents dans un contexte industriel où le nombre de documents est élevé et le nombre de classes n'est pas connu *a priori*. La méthodologie présentée est la suivante. Dans un premier temps le nombre de classes de documents est estimé. Ensuite un partitionnement de l'ensemble des documents est créé. Le centre de chacun des regroupements est extrait et servira d' *image de référence*, c'est à dire d'image requête pour effectuer un tri par similarité de style CBIR ("Content Based Image Retrieval"). Les premiers tests ont montré qu'en moyenne, les bases d'images de documents sont labellisées 3 fois plus rapidement avec notre système d'aide à la classification qu'avec une classification manuelle standard.

Abstract – This article deals with document image classification problem in industrial context where the number of document is large and the number of classes is unknown. The methodology introduced is the following. At first, the number of classes is estimated. Then a clustering of all the documents is created. The centers of each cluster is extracted and will be used as *reference images*, i.e. as query images in order to sort other images by similarity like CBIR (Content Based Image Retrieval). First tests show that, on average, databases are labelled 3 times faster with our assited classification tool than with manual classification.

1 Introduction

Les travaux présentés dans cet article se placent dans un cadre industriel où plusieurs dizaines de milliers de documents de type ressources humaines sont quotidiennement numérisés et indexés manuellement. Les contraintes principales sont les suivantes : aucune image n'est préalablement labellisée, le nombre de classes composant le jeu de donnée n'est pas connu à l'avance, il ne doit pas y avoir d'erreur de labellisation des images. Nous nous intéressons dans cet article à la définition d'une méthodologie permettant de simplifier et d'accélérer l'indexation manuelle des documents. Comme souligné récemment dans [13], cette problématique représente un enjeu économique important pour les sociétés de numérisation.

Chen et Blostein [3] présentent un état de l'art sur la classification d'images de documents. On constate grâce à leur étude que la plupart des méthodes de classification de documents sont des méthodes de classification supervisée. Par exemple, Shin *et al*. [14] proposent une approche basée sur les arbres de décision. En s'appuyant sur un ensemble de descripteurs pertinents, un arbre de décision est construit à partir de la vérité terrain extraite d'un échantillon représentatif de la base d'images. Cet arbre est ensuite utilisé pour classer le reste de la base. Sur une base de formulaires de 5590 images composée de 20 classes, en utilisant une vérité terrain de 2000 images, les auteurs obtiennent une précision de 99.7% sur l'indexation de 2000 autres images de la base. Dans [2], les auteurs proposent une approche basée sur les réseaux de neurones. La base de test est composée de 600 formulaires appartenant à 5 classes. La vérité terrain est composée de 305 images. Une précision de 92% est obtenue. Ces deux exemples montrent qu'il est nécessaire à la fois de connaître le nombre de classes et d'autre part de disposer d'une vérité terrain représentative de la distribution de la base réelle afin de procéder à la classification supervisée. Il faut également que la base d'apprentissage soit de taille suffisamment importante afin de permettre un apprentissage performant. Ces deux conditions sont d'autant plus difficiles à satisfaire que le volume d'images à indexer est important. De plus, malgré leurs bonnes performances, les résultats de ces techniques de classification doivent être vérifiés manuellement car ils n'atteignent pas une précision de 100%, ce qui est une contrainte industrielle.

Les auteurs de [1] proposent une méthode de classification nécessitant un nombre réduit de données d'apprentissage. Ils présentent deux variantes de K-Means utilisant des données labellisées pour chacune des classes. La première permet de réaliser l'initialisation aléatoire des K-Means à partir de données labellisées. La seconde permet de contraindre le processus de classification à ne pas changer de classe les éléments labellisés. Même si cette méthode nécessite un faible nombre de données d'apprentissages, il reste néanmoins complexe de trouver à l'avance le nombre de classes composant la base.

Les auteurs de [7] proposent également une méthode permettant d'intégrer des données labellisées. Cette méthode se base sur un apprentissage par contraintes binaires. Le principe consiste à montrer à un utilisateur deux images et à lui deman-

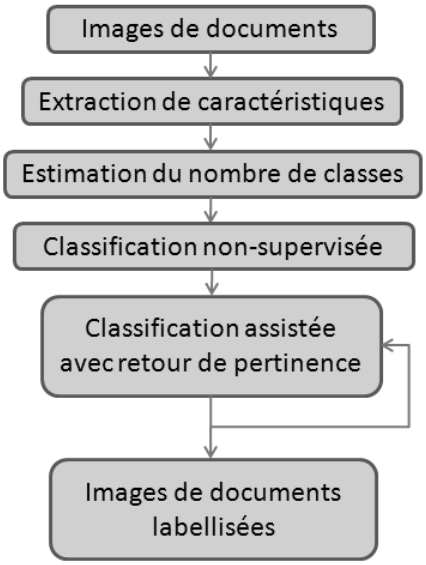

FIGURE 1 – Module d'aide à la classification. 1) Les caractéristiques sont extraites. 2) Le nombre de classes est estimé. 3) Un partitionnement des données est effectué. 4) Les documents sont classés à l'aide de technique CBIR avec retour de pertinence.

der si elles font parties de la même classe ou non. A partir de sa réponse une contrainte binaire attribuée à la paire d'images. Cette contrainte est soit "must link" soit "cannot link". Les tests montrent que le pourcentage d'images correctement labellisés tend vers 95% en définissant 30 contraintes binaires sur les 187 images de la base.

Pour remédier à la problématique précédemment citée, nous proposons dans cet article la méthodologie suivante. Dans un premier temps les descripteurs globaux (ceux de [14]) sont extraits. Sur la base de ces descripteurs, le nombre de classes k composant un jeu de données est estimé en utilisant le critère de silhouette [12]. Une fois k estimé, l'algorithme de classification non-supervisé PAM [9] permet d'obtenir un partitionnement en k classes avec pour chaque classe, un représentant pertinent qui sera appelé *image de référence*. Concrètement ce sont les médoïdes de PAM.

Un aspect intéressant de notre méthode est la mise en place d'un module d'aide à la classification manuelle basé sur les techniques de CBIR ("Content Based Image Retrieval"). Les images requêtes présentées à l'utilisateur sont les représentants extraits automatiquement par l'étape précédente. Notre proposition permet de regrouper puis de montrer à l'utilisateur des images possédant de fortes similarités visuelles. Le tri manuel est ainsi rendu plus simple et plus rapide. Un autre élément pour améliorer la mesure de similarité au sein d'une même classe est d'affiner les descripteurs utilisés afin de comparer et trier les documents au fur et à mesure des interventions de l'utilisateur. Il est ainsi possible d'utiliser les méthodes de retour de pertinence [16]. L'ensemble de la méthodologie est résumé sur la figure 1.

Les premiers tests réalisés montrent qu'une base d'images est labellisée en moyenne $3,5$ fois plus rapidement avec l'outil d'aide à la classification que nous proposons plutôt qu'avec une classification manuelle standard.

2 Estimation du nombre de classes k

Lors de la première création d'une base d'images, le nombre de classes n'est pas nécessairement connu *a priori*. Une première étape consiste à estimer le nombre de classes présentes dans la base. Pour cela, différentes techniques existent. Une des plus courantes est basée sur les critères d'informations tels que BIC (Bayesian Information Criterion) ou AIC (Aikake Information Criterion). Une comparaison de ces critères est réalisée dans [10]. L'étude de la qualité de partitionnement peut également être utile afin de choisir un partitionnement parmi plusieurs. Les auteurs de [8] présentent des mesures de qualité de partitionnement. L'homogénéité (la distance intra-classe) ainsi que la séparation (la distance inter-classe) sont les mesures de qualité les plus couramment utilisées.

Le critère de la silhouette moyenne décrit dans [12] et [11] est une mesure pertinente pour évaluer la qualité d'un partitionnement. (Nous avons choisi d'utiliser ce critère car c'est une mesure concrète prenant en compte à la fois l'homogénéité et la séparation des différentes classes.) La silhouette d'un élément x (dans notre cas, représenté par un vecteur de caractéristique) est calculée à partir de la moyenne des distances entre l'élément x et le reste des éléments $a(x)$ de sa classe C_x. Le minimum des distances moyennes entre l'élément x et les autres centres de classes $b(x)$ est ensuite calculé. La silhouette de l'élément x se calcule de la manière suivante : $silh(x) = \frac{b(x)-a(x)}{\max(a(x),b(x))}$. Une fois la silhouette calculée sur chaque élément, il est possible de calculer la moyenne des silhouettes pour une classe : $S_{C_i} = \frac{\sum_{j \in C_i} silh(j)}{Card(C_i)}$. Finalement, la moyenne des silhouettes des classes est calculée $GS = \frac{\sum_{i=1}^{k} S_{C_i}}{k}$. Plus la valeur de GS est proche de 1 plus la partition est de bonne qualité car elle traduit une forte variabilité inter-classes et faible variabilité intra-classe. Afin de sélectionner le nombre de classes qui optimise le partitionnement, GS est calculée pour chaque valeur de k allant de 3 à K où K est spécifié par l'utilisateur (pour les tests, K a été fixé à $\sqrt{(tailleBase/2)}$). Le k retenu est celui qui maximise GS. Des tests ont été effectués sur 4 bases issues de productions industrielles. Les bases 1 et 3 sont composées uniquement de factures, les bases 2 et 4 sont composées de documents RH divers tel que des contrats de travail, des conventions de stage, des déclarations d'embauche, des fiches d'entretien personnel, de renseignement administratif, etc. Le tableau 1 illustre la pertinence du critère de la silhouette pour estimer automatiquement le nombre de classes composant une base d'images.

Une fois que le nombre de classes est estimé, l'ensemble des données est partitionné à l'aide d'un algorithme de classification non-supervisé. Pour cela nous avons choisi la méthode de partitionnement PAM (Partitioning Around Medoids). Il est a

TABLE 1 – Synthèse des tests réalisés sur l'estimation du nombre de classes à l'aide de la mesure de silhouette. Pour les bases 1 et 3, le k estimé est très proche du k réel. Pour les bases 2 et 4, le k est sous-évalué. Une étude approfondie de ces deux bases montre que les documents composant certaines classes sont visuellement similaires et que certaines des classes sont marginales car elles contiennent peu d'éléments.

base	images	k "réel"	k silh
BD1	1509	7	6
BD2	883	19	13
BD3	2574	33	35
BD4	3352	30	13

noté également que certaines techniques de classification non supervisées combinent la phase d'estimation du nombre de classes et la phase de partitionnement dans un même algorithme. On peut citer notamment le competitive clustering [6] ou Mclust [5].

3 Classification assistée avec retour de pertinence

L'estimation automatique du nombre de classes k permet, à partir des médoïdes de chacune des k classes, d'extraire k images de référence représentant au mieux chacune des classes. La classification assistée est réalisée en montrant successivement à l'utilisateur une des images de référence accompagnée des 50 images qui lui sont le plus similaires. L'utilisateur indique alors quelles sont les images qui n'appartiennent pas à la même classe que l'image présentée, une nouvelle série de 50 images est alors proposée pour la même classe. Durant le traitement d'une classe, si parmi les 50 images affichées 25 images ou plus sont indiquées comme n'appartenant pas à la classe, un algorithme de sélection de caractéristiques est exécuté. La figure 2 illustre ce principe.

A partir des indications de l'utilisateur, un retour de pertinence peut être utilisé afin d'améliorer la mesure de similarité entre les documents. Ce retour de pertinence permet de réaliser une sélection de caractéristiques. Cette étape est importante puisque la nature des documents n'est pas connue à l'avance. Il est donc nécessaire d'extraire de nombreuses caractéristiques différentes puis de ne garder que celles qui sont pertinentes. L'algorithme de sélection de caractéristiques Fealect [1] permet de pondérer les dimensions du vecteur de descripteurs des images. Pour affecter ces poids, cet algorithme de sélection commence par générer plusieurs sous-ensembles d'éléments. L'algorithme détermine ensuite l'apport de chaque caractéristique dans le processus de classification à l'aide d'une régression des moindres angles (Least Angle Regression [4]).

Le tableau 2 récapitule les résultats obtenus lors de la classification de 4 bases d'images de documents de type "ressources humaines". L'utilisation de l'algorithme de sélection des meilleures caractéristiques permet un gain systématique sur le pourcentage de la base labellisée. Les tests réalisés en production sur plusieurs milliers d'images et plusieurs opérateurs, ont permis de montrer qu'en moyenne un opérateur professionnel met environ 8 secondes pour labelliser une image extraite aléatoirement de la base. Lorsqu'un opérateur utilise notre logiciel pour labelliser les images, il lui faut en moyenne 25 secondes pour sélectionner les images similaires à une image requête parmi les 50 images qui lui sont proposées.

FIGURE 2 – Classification assistée. Les documents sont triés par similarité décroissante. La première image encadrée en gris est l' *image de référence*. Les deux images encadrées en noir ont été sélectionnées par l'utilisateur car elle n'appartiennent pas à la même classe.

TABLE 2 – Résultats de la classification assistée (notée CA) avec retour de pertinence (notée RP). Le retour de pertinence permet de labelliser en moyenne au moins 2% supplémentaires de chaque base. La classification assistée avec retour de pertinence permet en moyenne, de diviser le temps d'indexation (notée TI) par $3,5$ par rapport à une indexation manuelle (notée CM).

		% de base labellisée		TI (minutes)	
base	images	sans RP	avec RP	CM	CA
BD1	1509	95,62624	99,3373	203,7	18,6
BD2	883	84,93771	86,97622	119,2	40,4
BD3	2574	89,66589	90,63714	347,5	101,7
BD4	3352	86,72434	88,30549	452,5	160,0

4 Conclusion et perspectives

Cet article présente une nouvelle méthodologie de classification d'images de documents de type ressources humaines. Le premier point important de notre proposition consiste en l'estimation du nombre de catégories de documents composant un

1. http ://cran.r-project.org/web/packages/FeaLect/FeaLect.pdf

jeu de données, là où les méthodes de l'état de l'art considèrent que cette information est donnée par l'utilisateur. La grande nouveauté de notre proposition réside dans la mise en place un système d'indexation d'images de documents basé sur la "requête par l'exemple" de style CBIR, dans lequel l'exemple est extrait automatiquement : se sont les *images de référence*, les médoïdes du partitionnement obtenus grâce à PAM. De plus, l'opérateur humain reste au cœur du système, garantissant que les images sont correctement labellisées. Au fur et à mesure que l'utilisateur classe les images de la base, notre système se spécialise permettant ainsi à l'opérateur humain d'indexer rapidement de grosses quantités d'images. Les tests effectués mettent en avant un gain de temps du processus d'indexation de l'ordre d'un facteur 3.

Les principales perspectives portent sur l'amélioration de l'étape de l'estimation du nombre k de classes images présentes dans la base. Pour cela, nous envisageons de combiner notre estimateur actuel avec d'autres estimateurs de qualité de partitionnement tels que le BIC [15]. L'utilisation de techniques semi-supervisée tel que [7] seraient envisageables car le nombre d'images à pré-labelliser est faible. Il serait également intéressant de comparer les différences entre les techniques de partitionnement de l'état de l'art (K-means, classification ascendante hiérarchique, etc.) afin de choisir une méthode de classification optimale. Enfin, une dernière piste d'amélioration des performances de la classification des images de documents réside dans la réduction du décalage sémantique (semantic gap) qu'il y a entre les attentes de l'utilisateur et les vecteurs de caractéristiques utilisés. Pour cela nous travaillons actuellement sur une adaptation de la méthodologie permettant à l'utilisateur de sélectionner des éléments de contenu (un logo, un tableau, etc.).

Remerciements

Nous tenons à remercier la société Gestform [2] pour son implication dans ces travaux. Gestform a fourni plus de 8000 images issues de ses productions. Elle a également fourni les statistiques relatives au temps de classement manuel des images.

Références

[1] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 27–34, 2002.

[2] F. Cesarini, M. Lastri, S. Marinai, and G. Soda. Encoding of modified XY trees for document classification. In *icdar*, page 1131. Published by the IEEE Computer Society, 2001.

[3] N. Chen and D. Blostein. A survey of document image classification : problem statement, classifier architecture and performance evaluation. *International Journal on Document Analysis and Recognition*, 10(1) :1–16, 2007.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2) :407–451, 2004.

[5] C. Fraley and A.E. Raftery. Enhanced model-based clustering, density estimation, and discriminant analysis software : MCLUST. *Journal of Classification*, 20(2) :263–286, 2003.

[6] H. Frigui and R. Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5) :450–465, 1999.

[7] N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering. *Pattern Recognition*, 41(5) :1834–1844, 2008.

[8] M. Halkidi, M. Vazirgiannis, and Y. Batistakis. Quality scheme assessment in the clustering process. *Principles of Data Mining and Knowledge Discovery*, pages 265–276, 2000.

[9] L. Kaufman and PJ Rousseeuw. Finding groups in data : an introduction to cluster analysis. *NY John Wiley & Sons*, 1990.

[10] A. Oliveira-Brochado and F.V. Martins. Assessing the number of components in mixture models : a review. *FEP Working Papers*, 2005.

[11] K. Pollard and M.J. Van Der Laan. A method to identify significant clusters in gene expression data. *Invited Proceedings of Sci2002*, 2 :318–325, 2002.

[12] P.J. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65, 1987.

[13] E. Saund. Scientific Challenges Underlying Production Document Processing. *Proceedings of Document Recognition and Retrieval XVIII*, 2011.

[14] C. Shin, D. Doermann, and A. Rosenfeld. Classification of document pages using structure-based features. *International Journal on Document Analysis and Recognition*, 3(4) :232–247, 2001.

[15] Q. Zhao, V. Hautamaki, and P. Franti. Knee Point Detection in BIC for Detecting the Number of Clusters. In *Advanced Concepts for Intelligent Vision Systems*, pages 664–673. Springer, 2008.

[16] X.S. Zhou and T.S. Huang. Relevance feedback in image retrieval : A comprehensive review. *Multimedia systems*, 8(6) :536–544, 2003.

2. http ://www.gestform.com