Catégorisation des méthodes de classification fondées sur l'Analyse de Concepts Formels

Hayfa Azibi, Nida Meddouri, Mondher Maddouri

Catégorisation des méthodes de classification fondées sur l'Analyse de Concepts Formels

Hayfa Azibi[1], Nida Meddouri[1,2], Mondher Maddouri[1,3]

[1] LIPAH, Faculté des Sciences de Tunis, Université El-Manar , Tunisie
hayfa.azibi@fst.utm.tn

[2] GREYC-CNRS UMR 6072, Université Caen Normandie, France
nida.meddouri@unicaen.fr

[3] CoB, Université de Jeddah, Arabie Saoudite
maddourimondher@yahoo.fr

Résumé : Les deux dernières décennies ont vu le développement de plusieurs méthodes de classification basées sur l'analyse de concepts formels (ACF). Dans cet article, nous présentons trois catégories de méthodes de classification basées sur l'ACF : des approches basées sur un classifieur unique, des approches basées sur une combinaison de classifieurs générés par une méthode d'ensemble et des approches basées sur un classifieur distribué.
Mots-clés : Intelligence artificielle, Fouille de données, Apprentissage automatique, Classification supervisée, Analyse de concepts formels, Méthodes d'ensemble, Big data.

1 Introduction

L'explosion du volume et la rapidité de la croissance des données ont introduit plusieurs défis dans de nombreux problèmes d'apprentissage du monde réel. La classification supervisée est une tâche de l'apprentissage automatique. L'objectif d'un problème de classification est de déterminer la classe avec laquelle seront étiquetées les nouvelles données.

L'analyse de concepts formels (ACF) (Ganter & Wille, 1999) est largement utilisée dans l'apprentissage automatique. L'ACF est une théorie mathématique basée sur les hiérarchies d'un treillis de concepts formels. C'est un cadre théorique qui structure un ensemble d'objet (appelé extension) et un ensemble d'attributs (appelé intention). L'extension couvre tous les objets appartenant au concept. L'intention est l'ensemble d'attributs qui caractérisent un objets.

Une approche de classification supervisée se fait en deux phases : une phase d'apprentissage et une phase de classement. Dans la phase d'apprentissage, un classifieur est généré à partir d'un modèle de classification ; en analysant des objets qui sont décrits par des attributs dans l'ensemble de données d'apprentissage. Chaque objet est censé appartenir à une classe prédéfinie et représentée par une étiquette précise dans l'ensemble de données d'apprentissage. Dans la phase de classement, le modèle construit précédemment est utilisé pour classer/étiqueter les nouveaux objets.

Dans littérature, plusieurs études comparatives ont été réalisées concernant les méthodes de classification par l'ACF. (Fu *et al.*, 2004) ont réalisé une étude comparative théorique et expérimentale sur quelques méthodes de classification par l'ACF. D'autres méthodes sont présentées dans (Meddouri & Maddouri, 2008) et qui se basent sur un classifieur unique. Les auteurs ont présenté les méthodes de classification par l'ACF en évoquant les notions de treillis complet, de demi-treillis et de couverture. (Trabelsi *et al.*, 2016) ont présenté une taxonomie des méthodes de classification supervisée existantes. Cette taxonomie propose deux catégories : des méthodes exhaustives et des méthodes combinatoires. La première catégorie se caractérise par l'utilisation d'un seul classifieur. La deuxième catégorie contient les méthodes combinatoires qui exploitent les paradigmes d'apprentissage à partir d'ensembles de classifieurs générés séquentiellement ou parallèlement. Par conséquent, le travail présenté dans cet article est de mettre à jour une catégorisation des méthodes en introduisant, entre autres, une nouvelle catégorie fondée sur un classifieur distribué.

2 Méthodes de classification basées sur l'ACF

Nous présentons trois catégories de méthodes de classification par l'ACF. Cependant, la principale différence entre ces catégories réside dans la façon que le classifieur est généré. En fait, ces méthodes se reposent sur l'utilisation d'un classifieur unique, une combinaison de classifieurs générés par une méthode d'ensemble ou d'un classifieur distribué.

2.1 Les méthodes fondées sur un classifieur unique

Les méthodes fondées sur un classifieur unique s'appuient sur la génération d'un treillis de concepts. Le treillis de concepts est une structure mathématique regroupant l'ensemble des concepts formels d'un contexte d'apprentissage ; et qui sont hiérarchiquement organisées par des relations de sous-concept/super-concept. Nous allons présenter dans la suite les méthodes de classification basées sur un classifieur unique selon le mode de construction du treillis.

La génération d'un treillis complet consiste à ajouter des concepts formels à ce treillis et mettre à jour les liaisons hiérarchiques qui se trouvent entre eux. De nombreux algorithmes de classification par l'ACF qui construisent un treillis complet, ont été développés. Nous citons GRAND (Oosthuizen, 1996), RULEARNER (Sahami, 1995) et NAVIGALA (Visani *et al.*, 2011).

Un demi-treillis est une structure mathématique qui représente une partie du treillis de manière sélective. Le processus de classification est le même pour les méthodes citées précédemment. Mais la principale différence entre elles est le nombre de concepts formels généré et à partir de quel demi-treillis (supérieur ou inférieur). Des méthodes comme LEGAL (Nguifo & Njiwoua, 2005) et CLANN (Tsopzé *et al.*, 2007) construisent un demi-treillis supérieur en réduisant considérablement leurs complexités théoriques et leurs temps d'exécution.

Une couverture de concepts est définie comme étant une partie du treillis qui ne contient que quelques concepts générés. IPR (Maddouri, 2004) et CITREC (Douar *et al.*, 2008) sont deux méthodes qui génèrent une couverture de concepts. IPR permet de générer une couverture de concepts pertinents. Cependant, IPR peut induire des concepts redondants. CITREC propose de réduire le contexte d'apprentissage et dans la suite générer un treillis complet à partir de ce contexte réduit. Une représentation condensée de données peut causer une perte d'information pour CITREC.

L'inconvénient majeur des méthodes citées précédemment demeure dans l'utilisation d'un seul classifieur, en outre une complexité importante et le type de données traitées qui sont binaires pour la plupart des méthodes. En conséquence, de nombreuses recherches dans la littérature se sont orientées vers la combinaison de méthodes de classification basées sur les méthodes d'ensemble par Boosting ou Bagging.

2.2 Les méthodes fondées sur les ensembles de classifieurs

Les méthodes d'ensemble Boosting (Freund, 1995) et Bagging (Breiman, 1996), sont des modèles d'apprentissage qui combinent les sorties de plusieurs classifieurs pour améliorer les performances. Le principe de Boosting fait référence à la combinaison d'un ensemble de classifieurs à travers un processus en cascade pour améliorer les décisions de classification du modèle produit. En revanche, le Bagging consiste à sous-échantillonner l'ensemble des données d'apprentissages et générer un classifieur pour chaque sous-échantillon. Il existe deux catégories de méthodes de classifications ensemblistes : les méthodes fondées sur le Boosting comme BFC (Meddouri & Maddouri, 2009) et BNC (Meddouri & Maddouri, 2010) et les méthodes fondées sur le Bagging comme DNC (Meddouri *et al.*, 2014) et B-RCL (Ali, 2018).

2.3 Les méthodes fondées sur un classifieur distribué

Au cours des dernières décennies, le volume de données générées à partir de diverses sources n'a cessé d'exploser. En effet, les algorithmes existants ne sont pas extensibles aux

nouveaux ensembles de données énormes pour l'extraction et la représentation des connaissances. Une nouvelle catégorie de méthodes de classification par l'ACF, comme Dist-CNC (Fray *et al.*, 2019), est proposée. Cette méthode permet l'extraction des connaissances à partir de grand volume de données dans un environnement distribué (cloud computing).

3 Discussion

La table 1 montre une comparaison des méthodes de classification fondées sur un classifieur unique à savoir : GRAND, LEGAL et IPR. Ces méthodes génèrent respectivement un treillis complet, un demi-treillis et une couverture de concepts à partir des données binaires. Ces méthodes traitent des données multiclasses à l'exception de LEGAL qui se limite à deux classes. Pour la construction de treillis, ces méthodes utilisent des algorithmes pour générer les treillis de concepts ; et qui peuvent être incrémentaux ou non-incrémentaux. Ces méthodes choisissent de représenter les connaissances apprises par des concepts pertinents ou des règles. Dans la phase de classement, chaque méthode utilise sa stratégie appropriée afin de prédire une classe pour chaque nouvel objet.

TABLE *1 – Comparaison des méthodes de classification basées sur un classifieur unique*

Méthode	**GRAND**	**LEGAL**	**IPR**
Type de données	Binaire	Binaire	Binaire
Numbre de classes	Multi-classe	2 classes	Multi-classe
Strusture de concepts	Treillis complet	Demi-treillis	Couverture
Algorithme de construction de treillis	Ossthuizen	Bordat	Approche heuristique
Incrémental	Oui	Non	Oui
Sélection de concepts	Cohérence maximalité	Cohérence maximalité	Entropie de Shanoon
Connaissance apprise	Règles	Concepts pertinents	Règles
Classification	Vote	Vote	Règles pondérées
Complexité	O($2^l \times l^4$ avec l=min(n,m)	O(($\|L\|$ × n (1-α)) avec $\|L\|$ = nombre de concepts, α = critère de validité	O($n^2 \times m^2 \times$ (m+n))

La table 2 présente une comparaison des méthodes de classification fondées sur les ensembles de classifieurs. Les méthodes présentées varient en fonction de l'approche d'apprentissage : séquentiel ou parallèle.

TABLE *2 – Comparison des méthodes de classification basées sur les méthodes d'ensemble*

Méthode	**BFC**	**BNC**	**DNC**	**B-RCL**
Structure de concepts	Couverture	Couverture	Couverture	Demi-treillis
Type de données	Binaire	Nominal	Nominal	Nominal
Sélection de concept	Entropie	Gain informationnel	Gain informationnel	Couverture conceptuelle aléatoire
Connaissance apprise	Règle	Règles	Règles	Règles
Classification	Vote pondéré	Vote pondéré	Vote majoritaire	Vote majoritaire
Ensemble	Séquentiel	Séquentiel	Parallèle	Parallèle
Complexité	O(nlog(n)+ nm)	O(nlog(n)+ nm') avec m' attributs nominaux	O(n') avec n' taille du sous-échantillon stratifié	O(N^3) avec N est le nombre de classifieurs

Les tables 1 et 2 montrent également une comparaison des complexités théoriques où *n* est le nombre d'objets et *m* est le nombre d'attributs. GRAND a une complexité exponentielle, car il navigue dans la totalité de l'espace de recherche (le treillis de concepts). LEGAL construit un demi-treillis ce qui réduit considérablement cette complexité. IPR a la complexité minimale parmi ces méthodes grâce à la génération des concepts les plus pertinents. Comme l'illustre le tableau 2, les méthodes basées sur l'apprentissage parallèle comme DNC et B-RCL atteignent respectivement une complexité linéaire et une complexité polynomiale. L'extraction de connaissances à partir de grands ensembles de données reste un défi et une tâche difficiles pour l'outil traditionnel d'exploration de données. Les classifieurs distribués deviennent une solution pour répondre à ce problème.

4 Conclusion

Dans cet article, nous avons présenté trois catégories de méthodes de classification fondées sur l'ACF. Ces méthodes se divisent en méthodes fondées sur un classifieur unique, méthodes fondées sur les ensembles de classifieurs et une nouvelle catégorie de méthodes fondées sur un classifieur distribué.

Références

ALI M. A. T. (2018). Bagged randomized conceptual machine learning method. Master's thesis, College of Engineering, Qatar.

BREIMAN L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

DOUAR B., LATIRI C. & SLIMANI Y. (2008). Approche hybride de classification supervisée à base de treillis de galois : application à la reconnaissance de visages. In *Actes des 8èmes Journées Francophones en Extraction et Gestion des Connaissances*, volume E-11 of *Revue des Nouvelles Technologies de l'Information*, p. 309–320 : Cépaduès-Éditions.

FRAY R., MEDDOURI N. & MADDOURI M. (2019). Cloud implementation of classier nominal concepts using distributedwekaspark. In *Supplementary Proceedings of ICFCA 2019 Conference and Workshops*, volume 2378 of *CEUR Workshop Proceedings*, p. 125–136 : CEUR-WS.org.

FREUND Y. (1995). Boosting a weak learning algorithm by majority. *Information and computation*, **121**(2), 256–285.

FU H., FU H., NJIWOUA P. & NGUIFO E. M. (2004). A comparative study of fca-based supervised classification algorithms. In *Proceeding of Second International Conference on Formal Concept Analysis*, p. 313–320.

GANTER B. & WILLE R. (1999). Formal concept analysis, mathematical foundation. Springer.

MADDOURI M. (2004). Towards a machine learning approach based on incremental concept formation. *Journal of Intelligent Data Analysis*, **8**(3), 267–280.

MEDDOURI N., KHOUFI H. & MADDOURI M. (2014). Parallel learning and classification for rules based on formal concepts. In *Proceedings of the 18th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Procedia Computer Science, p. 358–367 : Elsevier.

MEDDOURI N. & MADDOURI M. (2008). Classification methods based on formal concept analysis. In *Proceedings of the 6th International Conference on Concept Lattices and Their Applications*, p. 9–16.

MEDDOURI N. & MADDOURI M. (2009). Boosting formal concepts to discover classification rules. In *Proceeding of the 22rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, volume 5579 of *Lecture Notes in Computer Science*, p. 501–510 : Springer.

MEDDOURI N. & MADDOURI M. (2010). Adaptive learning of nominal concepts for supervised classification. In *Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6276 of *Lecture Notes in Computer Science*, p. 121–130 : Springer.

NGUIFO E. M. & NJIWOUA P. (2005). Treillis de concepts et classification supervisée. *Journal of Technique et Science Informatiques*, **24**(4), 449–488.

OOSTHUIZEN G. (1996). The application of concept lattice to machine learning. *Dept. Comput. Sci., Univ. Pretoria, Pretoria, South Africa, Tech. Rep. CSTR*, **94**(01).

SAHAMI M. (1995). Learning classification rules using lattices (extended abstract). In *Proceedings of the 8th European Conference on Machine Learning*, volume 912 of *Lecture Notes in Computer Science*, p. 343–346 : Springer.

TRABELSI M., MEDDOURI N. & MADDOURI M. (2016). New taxonomy of classification methods based on formal concepts analysis. In *Proceedings of the 5th International Workshop "What can FCA do for Artificial Intelligence" ? co-located with the European Conference on Artificial Intelligence*, volume 1703, p. 113–120.

TSOPZÉ N., MEPHU NGUIFO E. & TINDO G. (2007). Clann : Concept lattice-based artificial neural network for supervised classification. In *Proceedings of the 5th International Conference on Concept Lattices and Their Applications*, volume 331.

VISANI M., BERTET K. & OGIER J. (2011). Navigala : an original symbol classifier based on navigation through a galois lattice. *International Journal of Pattern Recognition and Artificial Intelligence*, **25**(4), 449–473.