Classification multi-labels de données de santé médico-économiques

Max Chevalier, Thierno Ibrahima Diop, Imen Megdiche, Nathalie Bricon-Souf, Olivier Teste

Classification multi-labels de données de santé médico-économiques

Max CHEVALIER[(*,**)], **Thierno Ibrahima DIOP** [(*,***)], **Imen MEGDICHE**[(*,**)], **Nathalie SOUF**[(*,**)], **Olivier TESTE**[(*,**)]
max.chevalier@irit.fr, thierno.diop@esp.sn, imen.megdiche@irit.fr, nathalie.souf@irit.fr, olivier.teste@irit.fr

[(*)] IRIT, Institut de Recherche en Informatique de Toulouse, UMR 5505, FRANCE
[(**)] Université de Toulouse, FRANCE
[(***)] Ecole Supérieure Polytechnique, Dakar, SENEGAL

Mots clefs :
Systèmes d'informations, Qualité des données, Apprentissage automatique, Classification multi-labels, Analyse Prédictive, Données de santé médico-économiques, PMSI, Expérimentations, Algorithmes.
Keywords:
Information Systems, Data quality, Machine learning, Multilabel classification, predictive analysis, Medico-economic health data, PMSI, experimentations, algorithms.

Palabras clave :
sistemas de información, calidad de datos, Aprendizaje automático,Clasificación multicategoría, análisis predictivo, datos de salud medicoeconómica, PMSI, experimentos, algoritmos.

Résumé

Cet article présente une application de la classification multi-labels pour la prédiction des diagnostics secondaires à partir d'un diagnostic primaire connu (et d'informations complémentaires) dans les données du PMSI (Programme de Médicalisation des Systèmes d'Information). Nous commençons, dans cet article, par exposer le contexte de ces travaux et justifier l'application d'une classification multi-labels pour répondre à ce besoin. Une étude bibliographique sur la classification multi-labels est ensuite présentée en mettant l'accent sur les caractéristiques des données multi-labels que nous manipulons et les différentes métriques utilisées dans ce domaine pour évaluer la qualité des jeux de données. Nous menons ensuite une étude comparative des algorithmes de la littérature appliqués au contexte de la prédiction de diagnostics afin d'identifier la meilleure stratégie à appliquer. Cette étude est composée de trois étapes: (i) caractérisation de la qualité des données multi-labels, (ii) relevé des expérimentations menées sur les différents algorithmes et (iii) comparaison des résultats en s'appuyant sur les métriques de référence.

1 Introduction

Le PMSI[1] (Programme de Médicalisation des Systèmes d'Information) permet de rendre compte de façon standardisée de l'activité des hôpitaux en mémorisant des informations permettant d'évaluer l'activité pour chaque séjour de patient. Il contient notamment les informations relatives aux diagnostics et celles relatives aux actes médicaux effectués. Ces éléments sont codés au moyen de classifications médicales (respectivement CIM10- Classification Internationale des Maladies 10ème version et CCAM - Classification Commune des Actes Médicaux) permettant ainsi une représentation standardisée de ces informations médicales. La tarification à l'activité (T2A) qui permet à l'heure actuelle d'établir le montant des revenus des hôpitaux, est ainsi directement liée au codage des diagnostics. La connaissance fine des diagnostics pour les séjours est très importante d'une part pour avoir une base d'information de qualité dans une perspective d'amélioration de soins, d'autre part pour permettre un financement des hôpitaux qui corresponde au mieux aux soins effectués. Les hôpitaux doivent donc renseigner les diagnostics des patients en les codant ; les codeurs se basent sur l'ensemble des documents médicaux (résultats d'IRM, prescriptions médicales, etc.) associés au patient et le référentiel complexe que constitue la CIM, contenant plus de 33 000 codes représentant les différents diagnostics. Cette tâche est complexe. Chaque situation donnant lieu à un diagnostic principal mais également à d'éventuels autres diagnostics, certains d'entre eux peuvent parfois être oubliés ou mal codés lors du processus de codage.

L'objectif de ce travail est de proposer une aide à l'activité de codage. Cette aide consiste à utiliser des prédictions de diagnostics obtenues à partir d'algorithmes d'apprentissage automatique ; les algorithmes exploitent les diagnostics secondaires reliés au diagnostic primaire. Cette proposition a pour ambition d'accélérer et faciliter la saisie des diagnostics des patients par les codeurs dans les hôpitaux.

Notre travail se focalise sur la prédiction des diagnostics secondaires, qui sont les plus difficiles à identifier. Ce travail est réalisé en partant de la base de données du PMSI à une échelle locale. Cette base nous est rendue accessible, sous clauses de confidentialité, par le département d'information médical du Centre Hospitalier Intercommunal de Castres-Mazamet (CHIC).

La problématique se définit donc ainsi : connaissant, pour un séjour donné, des informations générales ainsi que le diagnostic primaire[2], comment prédire le ou les diagnostics secondaires[3] les plus probables d'être associés à ce séjour ?

Dans littérature, nous pouvons identifier certains travaux qui ont traité le sujet tels que [5] et [2]. Les auteurs de [5] appliquent des algorithmes de machine learning, en l'occurrence des arbres de décision et des approches Naïves Bayésien, pour prédire un seul diagnostic secondaire à la fois pour chaque diagnostic primaire. Les auteurs de [2] utilisent le « *Golay Code* » pour leur prédiction. Cependant, le motif du séjour (le diagnostic primaire) est souvent associé à plusieurs diagnostics secondaires. Il est donc intéressant de prolonger les travaux existants qui prédisent un diagnostic sachant le diagnostic primaire en essayant de nous rapprocher des caractéristiques des données médicales utilisées qui combinent plusieurs diagnostics pour un même séjour. Nous avons ainsi privilégié la piste de la classification multi-labels. Ce type de classification regroupe les algorithmes de classification pouvant donner en sortie plusieurs labels au lieu d'un seul. Ces algorithmes sont très appropriés à notre contexte étant donné que nous souhaitons prédire pour chaque diagnostic primaire un ensemble de diagnostics secondaires qui lui sont associés, chacun des diagnostics secondaires pouvant être vu comme un label. Après avoir identifié que la méthode de classification multi-labels sera appliquée, l'objectif est donc de trouver le meilleur algorithme pouvant prédire le mieux possible les diagnostics secondaires sur le jeu de données qui nous a été fourni.

---

[1] Programme de médicalisation des systèmes d'informations

[2] Motif principal du séjour du patient

[3] Diagnostics surgissant au cours du séjour du patient

2 Etude bibliographique sur la classification multi-labels

La classification dans le domaine de l'apprentissage automatique est une tâche prédictive qui permet d'apprendre sur des observations labellisées (c'est-à-dire « étiquetées ») afin de prédire les labels des nouvelles observations. Il existe plusieurs types de classifications en fonction du sujet à prédire [3] :

- La ***classification binaire*** : un seul label à prédire et ce dernier ne peut avoir que deux valeurs binaires (par exemple 0 ou 1, oui ou non, ...etc.) ;
- La ***classification multi-classes*** : un seul label à prédire, mais le label a plusieurs valeurs possibles (par exemple rouge, noire, bleu, ...etc.) ;
- La ***classification multi-labels*** : plusieurs labels à prédire et chaque label n'a que deux valeurs possibles ;
- La ***classification multi-dimensionnelle*** : plusieurs labels à prédire, mais au moins un des labels a plusieurs valeurs possibles ;
- La ***classification multi-instances*** : plusieurs instances ou observations peuvent être associées à un label.

La suite de ce document portera sur la classification multi-labels, ses concepts, les défis sous-jacents et les différentes approches possibles pour résoudre un tel type de problème de classification.

La classification multi-labels fait correspondre une entrée X (x_1, x_2, ..., x_I) à un vecteur binaire Y (y_1,y_2, …, y_N) en associant une valeur entre 0 et 1 à chaque élément (label) de ce vecteur. Le résultat d'un classificateur multi-labels peut être sous la forme d'une bipartition (labels pertinents et labels non pertinents) ou sous la forme d'un classement en mettant le label le plus pertinent en premier et en dernier le moins pertinent.

Les défis de la classification multi-labels

La classification multi-labels apporte en plus de la classification traditionnelle son lot de problèmes [3] à savoir :

- **Le défi de corrélation entre les labels** : la plupart des algorithmes proposés dans la littérature pour résoudre la classification multi-labels se basent sur des processus de simplification comme la transformation binaire et ne prennent donc pas en compte la corrélation souvent présente entre les labels au niveau du jeu de données, alors que cette information est susceptible d'aider à obtenir un meilleur modèle.
- **Le défi de grandes dimensions** : le fléau de la dimension est un obstacle majeur dans l'apprentissage automatique, car il augmente considérablement le temps d'apprentissage tout comme il diminue les performances des modèles. Il est encore plus présent dans la classification multi-labels car non seulement il y a les variables indépendantes mais il y a aussi les labels qui sont touchés par ce fléau. De plus, les algorithmes supervisés utilisés dans la classification traditionnelle ne peuvent pas être directement utilisés pour diminuer les variables indépendantes car ils ne prennent en compte qu'un seul label. De plus, il faut penser à de nouveaux algorithmes [3] pour diminuer le nombre de labels, car on ne peut pas éliminer des labels comme on le fait avec les attributs.
- **Le défi de déséquilibre des labels** : le déséquilibre au niveau des labels est un problème majeur dans la classification de manière générale car la plupart des algorithmes de classification ne gèrent pas bien les classes minoritaires. C'est un problème qui est donc présent dans la classification traditionnelle. Plusieurs algorithmes ont été proposés afin de résoudre ce problème, sauf que dans la classification multi-labels, ces algorithmes ne peuvent pas être utilisés, car ces derniers ne prennent pas en compte le fait qu'il peut exister plusieurs labels en sortie. Dans la classification multi-labels une combinaison de labels peut être rare, tout comme un label rare peut être associé à des labels plus fréquents. De plus, certaines approches comme la transformation binaire augmentent le niveau

de déséquilibre dans le jeu de données obtenu après transformation. Dans la classification multi-labels plusieurs algorithmes ont été proposés pour faire face à ce problème :

- Echantillonnage ;
- Adaptation des algorithmes de classification ;
- Apprentissage sensible au coût : une combinaison des deux méthodes ci-dessus.

Caractéristiques des jeux de données multi-labels

Avant d'essayer de résoudre un problème de classification multi-labels quelconque, il est pertinent de connaître les caractéristiques du jeu de données multi-labels afin d'en évaluer sa qualité. Plusieurs métriques ont été proposées dans la littérature pour définir ces caractéristiques :

- Métriques de base
  - ***Cardinalité des labels*** : cette mesure permet de connaître à quel point le jeu de données est multi-labellisé, c'est-à-dire combien de labels il y a par instance en moyenne ;
  - ***Densité des labels*** : normalise la cardinalité en la pénalisant par le nombre de labels ; en effet la cardinalité est influencée par le nombre de labels dans le jeu de données.
- Métriques de déséquilibre
  - ***IRLbl(l)*** : mesure le niveau de déséquilibre pour chaque label. Ceci permet de savoir à quel point un label est présent dans le jeu de données ;
  - ***MeanIR*** : permet de connaître le niveau de déséquilibre de manière globale ;
  - ***CVIR*** : permet de connaître ce qui a causé une valeur importante de la moyenne ci-dessus (*MeanIR*) ; une valeur proche de 0 permet d'affirmer que le déséquilibre est présent sur plusieurs labels et une valeur proche de 1 indique que le déséquilibre est surtout accentué sur une minorité de labels.
- Autres métriques
  - ***SCUMBLE*** : mesure la concurrence entre les labels fréquents et les labels rares. Une valeur importante indique qu'il sera difficile d'obtenir un bon modèle avec le jeu de données ;
  - ***TCS*** : mesure la complexité théorique du jeu de données. Une valeur importante indique un temps d'apprentissage long et un modèle complexe.

Les approches de résolution dans la classification multi-labels

Pour résoudre le problème de la classification multi-labels, trois approches sont possibles [3] :

- ***Approche par transformation*** : le problème est transformé en une classification binaire ou multi-classes en créant un nouveau jeu de données à partir du jeu de données d'origine afin que les algorithmes traditionnels puissent être utilisés pour obtenir un modèle ;
- ***Approche par adaptation*** : les algorithmes traditionnels sont modifiés et adaptés afin de pouvoir travailler avec des jeux de données multi-labels et donner en sortie plusieurs labels au lieu d'un seul ;
- ***Approche ensembliste*** : cette approche est une suite logique de la première et consiste à utiliser un ensemble de classificateurs afin de résoudre le problème de classification.

Mesures de performance d'un classificateur multi-labels

Contrairement à un classificateur traditionnel, un classificateur multi-labels peut faire une prédiction totalement correcte, partiellement correcte ou totalement incorrecte. C'est la raison pour laquelle on ne peut pas utiliser les mêmes métriques de performance que celles utilisées dans la classification traditionnelle. Les métriques de performance les plus utilisées dans la classification multi-labels peuvent être regroupées en quatre catégories [3, 7, 10] :

- **Métriques basées sur le temps** : elles regroupent les temps d'apprentissage et de test ;
- **Métriques basées sur les observations** : elles sont calculées séparément pour chaque observation. Leurs valeurs sont obtenues en calculant la moyenne.
  - ***Hamming Loss*** : la différence symétrique entre les labels faussement prédits et les labels réels ;
  - ***Accuracy***: la proportion entre les labels correctement prédits et les labels actifs (union des labels prédits et réels) ;
  - ***Precision***: la proportion entre les labels correctement prédits et les labels prédits ;
  - ***Recall*** : la proportion entre les labels correctement prédits et les labels réels ;
  - ***F1*** : la moyenne harmonique des deux dernières métriques (rappel et précision) ;
  - ***Exact Match*** : mesure la plus stricte, elle donne la proportion entre les observations correctement prédites (tous les labels ont été bien prédit) et le totale des observations.
- **Métriques basées sur les labels** : elles sont calculées indépendamment pour chaque label. Pour calculer la moyenne, deux approches sont possibles :
  - ***Moyenne Macro* : l**es métriques sont calculées individuellement pour chaque label et la moyenne est obtenue en les divisant sur le nombre de labels ;
  - ***Moyenne Micro* :** les prédictions correctes et fausses pour chaque label sont d'abord sommées, puis pour avoir la métrique en question (F1, Recall, Precision), on applique sa formule sur la somme obtenue ;
- **Métriques basées sur le classement** : ce sont des mesures pour apprécier les classificateurs multi-labels qui donnent en sortie un classement et non une bipartition :
  - ***Average precision***: le nombre de labels à parcourir avant de trouver un label non pertinent ;
  - ***Coverage***: le nombre de labels à parcourir pour trouver tous les labels pertinents ;
  - ***OneError*** : le nombre de labels en première position au niveau du classement et qui ne sont pas pertinents ;
  - ***RLoss*** : le nombre de fois où un label non pertinent est placé au-dessus d'un label pertinent.

3 Prédiction de diagnostics secondaires avec la classification multi-labels

Dans cette section, nous présentons la méthode proposée pour prédire les diagnostics secondaires à partir des diagnostics primaires sur la base PMSI. Notre méthode se compose de trois étapes : (1) exploration statistique des jeux de données initiaux et choix des algorithmes de classification multi-labels, (2) résolution du problème de déséquilibre des données et (3) comparaison des résultats des algorithmes et conclusion sur le meilleur candidat pour résoudre notre problème. L'environnement technique pour réaliser ce travail est composé de : R, MEKA, MULAN et SPARK.

Les diagnostics (primaires ou secondaires) sont encodés dans la base PMSI à l'aide de la CIM10. C'est une classification hiérarchique qui explicite chaque maladie avec un niveau de précision important comme le montre l'exemple présenté dans la Figure 1.

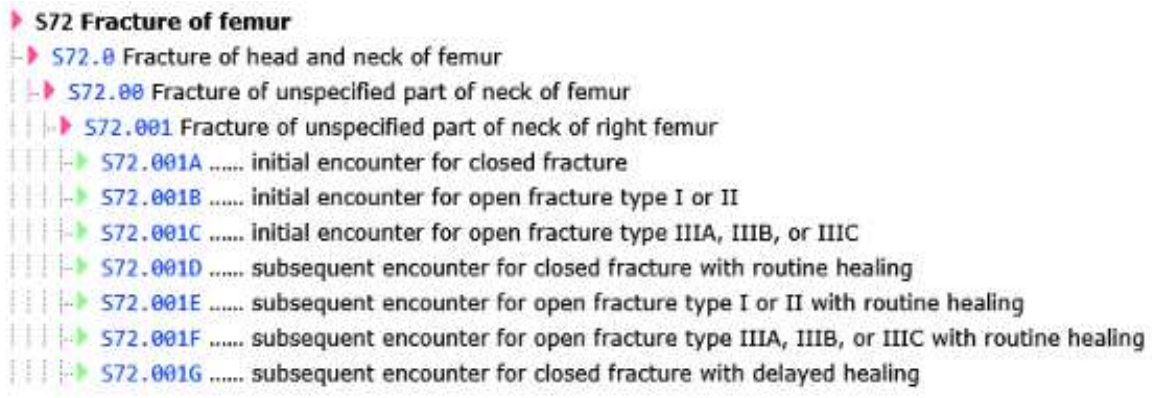

Figure 1. Exemple de code de diagnostic S72 - CIM 10

Les experts en codage ont affirmé que les trois premiers caractères du codage en CIM10 (le codage peut aller jusqu'à 8 caractères) apportent une information suffisante sur le diagnostic correspondant. Nous avons donc travaillé en prenant en compte la hiérarchie de codage effectuée au niveau des 3 premiers caractères de la CIM10 ce qui nous a permis de diminuer considérablement le nombre de labels (classes) sur le jeu de données (1705 au lieu de 7423 labels).

En outre nous avons décomposé le jeu de données initial en plusieurs jeux de données (i.e. un jeu de données pour chaque diagnostic primaire), ce qui a produit 1100 jeux de données au final. Vu le nombre élevé de jeux de données, seuls les 33 jeux de données les plus pertinents ont été choisis pour faire l'étude présentée dans cet article. Ce choix correspond aux diagnostics primaires associés aux 8 diagnostics secondaires jugés comme étant les plus difficiles à détecter par les experts [5].

Après l'analyse des caractéristiques des jeux de données, une première expérimentation avec différents algorithmes de classification multi-labels choisis en fonction des caractéristiques des données a été effectuée, et un déséquilibre au niveau des labels a été observé sur tous les jeux de donnée ; l'algorithme ML_ROS a été sélectionné et appliqué pour rééquilibrer ces derniers. Les mêmes algorithmes de classification ont été réutilisés sur les jeux de données obtenus après l'application de l'algorithme ML_ROS pour également apprécier l'impact du rééquilibrage sur la performance des prédictions.

3.1 Caractéristiques des jeux de données initiaux et choix des algorithmes

Une étude préalable a été faite sur les 33 jeux de données pour connaître leurs caractéristiques afin de mieux choisir les algorithmes de classification multi-labels à utiliser dans nos expérimentations.

	Combinaison de labels			Déséquilibrage		Concurrence entre les labels	
DP	N.CL	N.CL.U	Card	meanIR	CVIR	scumble	scumble.cv
I50	2200	2182	11.8869	859.0958	1.0033	0.6155	0.3538
K65	**38**	**38**	10.7105	**26.7134**	**0.4798**	**0.4057**	0.3157
R10	786	715	**3.9674**	616.2670	0.7436	0.5523	0.4932
R52	166	164	7.0476	103.7168	0.5928	0.5510	**0.2742**
S06	1686	1621	6.5815	872.0711	0.8677	0.6174	0.3114
Z43	164	153	4.7511	133.3941	0.6157	0.4469	0.6132
Z51	1815	1296	4.0101	7225.8966	1.0650	0.6118	0.3275

Figure 2, extrait des caractéristiques des jeux de données (en rouge ou en gras les valeurs extrêmes)

Après l'analyse statistique des caractéristiques des jeux de données (un extrait des caractéristiques de quelques jeux de données est fourni dans la figure 2), trois constats majeurs ont été observés :

- ***Concurrence entre les labels*** : la concurrence entre les labels fréquents et rares est très élevée, ceci complique l'apprentissage des algorithmes, car les patterns sont encore plus difficiles à identifier ;
- ***Déséquilibre au niveau des labels*** *:* on remarque un déséquilibre très important des labels sur l'ensemble des jeux de données, compte tenu de leur valeur *MeanIR* et *CVIR*. Les auteurs de [4] affirment qu'un jeu de données multi-labels est déséquilibré lorsque la valeur de MeanIR > 1,5 et celle de CVIR > 0,2. Or ces valeurs sont largement dépassées sur nos jeux de données. La valeur minimale de *MeanIR* est de 26,7134 et la valeur minimale de CVIR est 0.4057 ;
- ***Combinaisons des labels très rares (parfois même uniques)*** : les combinaisons de labels sur les différentes observations se répètent rarement voir jamais, ce qui rend l'apprentissage plus compliqué pour certains types d'algorithmes.

La rareté des combinaisons des labels indique clairement qu'une méthode qui se base sur la combinaison de labels pour l'apprentissage n'est pas adaptée dans ce contexte, car la plupart des combinaisons de labels sont uniques ou très peu fréquentes. C'est pourquoi des algorithmes qui se basent sur la pertinence binaire mais qui prennent aussi en compte les corrélations entre les labels ont été choisis pour mener les expérimentations afin de choisir le plus performant. Ce choix est aussi justifié par la présence des catégories des diagnostics qui permettent de détecter une certaine corrélation entre les labels. Cependant pour s'assurer de cette hypothèse l'algorithme ***RAkEL*** qui se base sur la combinaison de labels est aussi évalué dans l'expérimentation. L'algorithme ***BPNN*** qui utilise les réseaux de neurones est également évalué.

Les algorithmes **BR**, **CC**, **BRq**, **BPNN** et **RAkEL** ont donc été évalués dans les expérimentations que nous présentons dans la suite de l'article. Avant cela, nous discutons du problème de déséquilibre des données et présentons la procédure mise en place.

3.2 Rééquilibrage des jeux de données

Comme constaté lors de l'étude des caractéristiques des jeux de données, tous sans exception sont déséquilibrés, c'est pourquoi l'application d'un algorithme d'échantillonnage est jugée nécessaire pour pouvoir améliorer les prédictions des modèles obtenus avec les jeux de données initiaux.

Les algorithmes de sous-échantillonnage ne sont pas envisageables dans notre contexte, car les jeux de données ont peu d'observations et on risquerait de perdre des informations capitales pour l'apprentissage telles que les « patterns » présents dans les observations éliminées. Les algorithmes de sur-échantillonnage se basant sur les combinaisons de labels ne sont pas non plus optimisés dans ce contexte dans la mesure où la plupart des combinaisons de labels sont rares et l'algorithme considérait donc qu'il n'y a pas de déséquilibre, surtout sur les jeux de données où on a très peu d'observations.

Les algorithmes de sur-échantillonnage se basant uniquement sur les labels [4] en considérant les métriques ***IR***, ***MeanIR*** et ***CVIR*** apparaissent donc comme les plus adaptés dans notre contexte pour les raisons citées ci-dessus. Ceci est confirmé par les résultats obtenus sur les expérimentations avec les algorithmes à pertinence binaires. L'algorithme ***ML_ROS*** (Multi Label Random Over Sampling) a été implémenté et utilisé pour réaliser le sur-échantillonnage des différents jeux de données de notre collection de données.

Après le rééquilibrage, une nette amélioration sur les métriques a été observée sur presque l'ensemble des jeux de données. Par exemple, le jeu de données Z51 passe de 7225 à 5464 pour la métrique *MeanIR*.

3.3 Extrait des résultats des expérimentations

Les jeux de données obtenus après le sur-échantillonnage avec l'algorithme ***ML_ROS*** ont été utilisés pour l'apprentissage des algorithmes de classification multi-labels afin d'apprécier l'amélioration des prédictions par rapport aux jeux de données initiaux. Le rééquilibrage améliore nettement les performances de prédiction comparé aux jeux de données initiaux. La figure 3 montre les performances obtenues sur les jeux de données rééquilibrés.

Toutes les trois catégories de métriques (par observation, par labels et classement) ont été plus ou moins « boostées » :

- Par observation : amélioration de l'ordre de 7% en moyenne ;
- Par labels : amélioration de l'ordre de 10% en moyenne ;
- Par classement : amélioration de l'ordre de 13% en moyenne.

Le temps d'apprentissage et de test a un peu augmenté en raison du nombre d'observations ajoutées par l'algorithme de rééquilibrage. Le meilleur algorithme de manière générale est ***BRq***, et le plus mauvais est ***RAkEL*** pour les raisons prévisibles et expliquées dans la section 3.1.

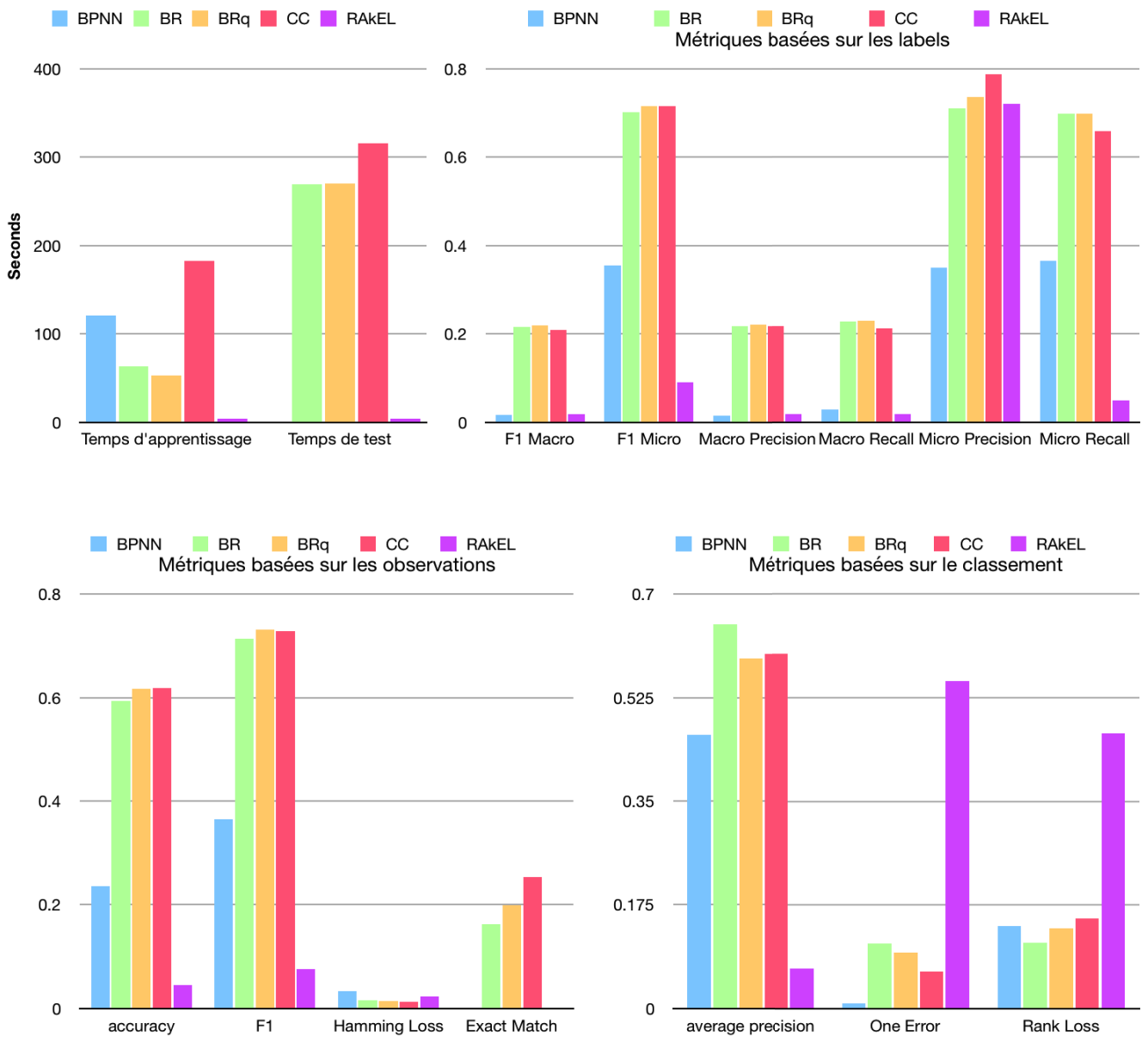

Figure 3, résultats des expérimentations sur les jeux de données rééquilibrés

4 Conclusion

L'étude de la classification multi-labels et les expérimentations que nous avons menées ont visé à mettre en évidence les caractéristiques des jeux de données manipulés, mais aussi ont permis de comparer différents algorithmes de classification multi-labels sur la base PMSI afin au final identifié celui qui était le plus adapté. Après l'analyse des caractéristiques des données, il a été constaté que les jeux de données manipulés dans les expérimentations sont déséquilibrés. Après analyse du déséquilibre, l'algorithme ***ML_ROS*** a été appliqué avec différents pourcentages afin d'obtenir une meilleure prédiction. Le sur-échantillonnage a nettement amélioré les prédictions, surtout avec ***ML_ROS25*** (résultat présenté sur ce document), c'est-à-dire ***ML_ROS*** avec un pourcentage égal à 25%.

Comme bilan de ce travail, nous concluons sur la pertinence du traitement du déséquilibre avec l'algorithme ***ML_ROS25*** sur les 1100 jeux de données ainsi que l'efficacité de l'algorithme ***BRq*** pour la classification multi-labels appliqué à la base PMSI puisqu'il s'agit de l'algorithme qui présente les meilleurs résultats si l'on considère l'ensemble des métriques.

Comme perspectives à ce travail, nous envisageons d'intégrer dans le processus d'analyse prédictive mise en place d'autres sources de données telles que les images médicales ou les données textuelles de santé afin d'améliorer encore la prédiction des diagnostics secondaires pour un séjour patient.

5 Bibliographie

[1] C. KOBY, D. MARK, G. KUZMAN, P. TALUKDAR PARTHA & C. STEVEN (2007). *Automatic code assignment to medical text*. Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing. 129-136. 10.3115/1572392.1572416.

[2] F. A. ALSABY. (2016). *Golay Code Classifier Approach for medical diagnosis.* SoutheastCon 2016, Norfolk, VA, 2016, pp. 1-6. doi: 10.1109/SECON.2016.7506676

[3] F. HERRERA, F. CHARTE, A.J. RIVERA, M.J. DEL JESUS (2016). *Multi-label Classification Problem Analysis. Metrics and Techniques*. Springer International Publishing, ISBN 78-3-319-41111-8.

[4] F. HERRERA, F.C HARTE, A.J. RIVERA, M.J. DEL JESUS (2015). *Addressing Imbalance in multi-label classification : Measures and random resampling algorithms*. Neurocomputing, Volume 163, 2015, Pages 3-16, ISSN 0925-2312.

[5] G. CHAHBANDARIAN, N.SOUF, I.MEGDICHE-BOUSARSAR, R.BASTIDE, J.C.STEINBACH (2017). *Predicting the encoding of secondary diagnoses. An experience based on decision trees.. Dans / In : Ingénierie des Systèmes d'Information, Hermès Science*, Vol. 22, N. 2, p. 69-94.

[6] H. HE, Y. MA (2013). *Imbalanced Learning : Foundations, Algorithms, and Applications*. Haibo He (Editor), Yunqian Ma (Editor), ISBN: 978-1-118-07462-6, Aug 2013, Wiley-IEEE Press

[7] H.GOUK, B.PFAHRINGER (2016). *Learning Distance Metric for multi-label* classification. JMLR: Workshop and Conference Proceedings 63:318–333, 2016

[8] S. T. MOTURU, H. LIU, W. G. JOHNSON (2008). *Understanding the Effects of Sampling on Healthcare Risk Modeling for the Prediction of Future High-Cost Patients*. In: FRED A., FILIPE J., GAMBOA H. (eds) Biomedical Engineering Systems and Technologies. BIOSTEC 2008. Communications in Computer and Information Science, vol 25. Springer, Berlin, Heidelberg

[9] R. MIOTTO, F. WANG, S. WANG, X. JIANG, JT DUDLEY (2017). *Deep learning for healthcare: review, opportunities and challenges*. Brief Bioinform. 2017 May 6. doi: 10.1093/bib/bbx044.

[10] T. GRIGORIOS & K. IOANNIS. (2009). *Multi-Label Classification : An Overview.* International Journal of Data Warehousing and Mining. 3. 1-13. 10.4018/jdwm.2007070101.