Predicting the encoding of secondary diagnoses

An experience based on decision trees

Ghazar Chahbandarian[1], **Nathalie Bricon-Souf**[1],
Imen Megdiche[1], **Rémi Bastide**[1], **Jean-Christophe Steinbach**[2]

1. *IRIT/ISIS, Université de Toulouse, CNRS,*
81100 Castres, France
{ghazar.chahbandarian,nathalie.souf,imen.megdiche,remi.bastide}@irit.fr

2. *Centre Hospitalier Intercommunal de Castres Mazamet,*
Department of Medical Information,
81100 Castres, France
jean-christophe.steinbach@chic-cm.fr

ABSTRACT. *In order to measure the medical activity, hospitals are required to manually encode diagnoses concerning an inpatient episode using the International Classification of Disease (ICD-10). This task is time consuming and requires substantial training for the staff. In this paper, we are proposing an approach able to speed up and facilitate the tedious manual task of coding patient information, especially while coding some secondary diagnoses that are not well described in the medical resources such as discharge letters and medical records. Our approach leverages data mining techniques, and specifically decision trees, in order to explore medical databases that encode such diagnoses knowledge. It uses the stored structured information (age, gender, diagnoses count, medical procedures, etc.) to build a decision tree which assigns the appropriate secondary diagnosis code into the corresponding inpatient episode. We have evaluated our approach on the PMSI database using fine and coarse levels of diagnoses granularity. Three types of experimentations have been performed using different techniques to balance datasets. The results show a significant variation in the evaluation scores between the different techniques for the same studied diagnoses. We highlight the efficiency of the random sampling techniques regardless of the type of diagnoses and the type of measure (F1-measure, recall and precision).*

RÉSUMÉ. *Afin de mesurer l'activité médicale, les hôpitaux sont tenus de coder manuellement des informations concernant les séjours des patients hospitalisés en utilisant la Classification Internationale des Maladies (CIM-10). Cette tâche est chronophage et nécessite une formation importante pour le personnel en particulier pour le codage des diagnostics associés (secondaires). Afin d'assister les personnels hospitaliers dans leur tâche, nous proposons une approche basée sur les techniques de fouille de données et plus précisément les arbres de décision qui permet de* *prédire le codage des diagnostics associés. Les arbres de décision sont construits à partir des données structurées de la base PMSI (âge, sexe, nombre de diagnostics et actes médicaux ...). Ces arbres de décision sont facilement exploitables par un non spécialiste en informatique tel qu'un médecin. Deux niveaux de granularité de diagnostic ont été exploités selon que l'on choisisse de représenter le diagnostic de façon très précise (fin niveau de granularité) ou en se contentant de garder une information plus générale (niveau de granularité plus grossier) correspondant aux catégories de diagnostics. Trois types d'expérimentations ont été réalisés selon différentes techniques d'équilibrage de dataset. Les résultats obtenus indiquent qu'il existe une variation significative des scores d'évaluation entre les différentes techniques pour les mêmes diagnostics étudiés. Nous mettons en évidence l'efficacité des techniques "random sampling" quels que soient le type de diagnostic et le type de mesure (F1-mesure, le rappel et la précision). Nos résultats montrent également l'efficacité d'utiliser le niveau fin de granularité de diagnostic quel que soit le diagnostic étudié.*

KEYWORDS: *data mining, machine learning, decision tree, PMSI, secondary Diagnoses, coding ICD-10.*

MOTS-CLÉS : *fouille de données, apprentissage automatique, arbre de décision, PMSI, diagnostiques secondaires, CIM-10.*

DOI:10.3166/ISI.22.2.69-94 © 2017 Lavoisier

1. Introduction

In France, since 1991, by recommendation of the ministry of Health, all public healthcare facilities are mandated to record patient diagnosis and medical procedures in a national database called PMSI (*Programme de Médicalisation des Systèmes d'Information*) equivalent to the PPS (*Prospective Payment System*) used in the USA (Fetter, 1991). The system was initially used for the purpose of reporting hospital activity and comparing the productivity between different facilities. In 1998, PMSI was used by all public and private hospitals for the purpose of fair funding. Since its creation, millions of records have been stored in the PMSI database, which makes it an attractive target for data analysis and prediction.

Each inpatient episode in France consists of one or several standard patient discharge reports called RUM (*Résumé Unité Médicale*). The RUM contains administrative information such as gender, age and length of stay. The RUM also contains medical information such as diagnoses and medical procedures performed during the stay in the medical unit. At the end of the inpatient episode, all the reports are combined into one report called RSS (*Résumé de Sortie Standardisé*). Then, an anonymisation process is applied in order to produce a so-called anonymised episode summary RSA (*Résumé de Sortie Anonymisé*). Finally, the RSA reports are sent to the Regional Health Agencies ARS (*Agences Régionales de Santé*) where they are stored in the national PMSI database. Each hospital is eventually refunded according to the activity described in the RSA reports.

Medical data is collected from different healthcare sources such as laboratory reports, radiology images, patient's consultations, observations and interpretations of the physician. Hospitals try to document their activities as accurately as possible to get fair payment. Inaccurate encodings of inpatient episode information could cause inaccurate refundings. Consequently, a lot of effort is made by hospitals to increase encoding accuracy of the diagnoses and medical procedures. Within each hospital the Medical Information Unit (*Département d'Information Médicale, DIM*) is responsible for the encoding process which is very sensible as explained by (Busse *et al.*, 2011) "If up-coding or incorrect coding is detected, hospitals must reimburse payments received. In addition, hospitals may have to pay high financial penalties of up to 5 per cent of their annual budgets".

Unlike primary diagnosis, which are not too difficult to encode, some secondary diagnoses require extra identification efforts. In fact, secondary diagnoses are often not clearly mentioned in the medical reports and cannot be directly deduced. In France, one hospital reported that more than one-third of patients having malnutrition and obesity as secondary diagnoses were not coded in the database (Potignon *et al.*, 2010). In order to identify all the secondary diagnoses, coders need to consider many sources and need to interpret information to find out the right code. Some form of support for semi-automatic code assignment can be a suitable solutions to speed up what coders have to do manually.

In this paper we are addressing the challenge of helping to automatically support the encoding of secondary diagnoses. In order to tackle this issue, we are proposing a new methodology based on data mining techniques, which are the best candidates to solve such prediction problems. Among the available methods in data mining, we focus on decision trees, which are extremely relevant since their results can be exploited by non-expert users in data mining field. Our methodology is applied on the PMSI national database. It is the richest and the most valuable source of documented standard diagnoses and medical procedures in France. In particular since it contains millions of records collected over the years and recently accessible for research purposes.

The rest of the paper is organized as follows: section 2 reviews some existing work within the diagnoses prediction domain; section 3 presents preliminary materials; section 4 details the proposed approach, while section 5 reports the experiments. Section 6 discusses the obtained results. The paper concludes in section 7 with some future perspectives.

2. Related work

In this section, we are surveying different research categories that have been proposed to predict diagnoses. Researchers address this prediction problem in a variety of applications such as marketing, e-business and other industrial sectors. Secondary diagnoses prediction and more generally data prediction in the healthcare domain have specific constraints since it is dealing with medical data, which is considered as unique in terms of heterogeneity, privacy-sensitive, ethical, legal, and social issues (J.Cios, Moore, 2002). Therefore, various methods are used to overcome these constraints and to solve the diagnosis encoding problem.

In the literature, encoding secondary diagnoses was performed through different techniques according to the different types of sources used. We can clearly distinguish two types of data sources used to predict diagnoses:

1. **Conventional data** where the main sources are clinical reports, physician's interpretations, discharge letters and other medical documents that are usually written in free text and that are frequently used by coders to determine the medical code.

2. **Structured data** where the main sources are the information stored in PMSI database, which contains well formated data concerning inpatient episodes.

(1) Conventional data

Natural Language Processing (NLP) methods are often used as a first step of data analysis. NLP consists of translating free text into formal representation or features so that machines can understand the text and manipulate it. Mining techniques can then be applied in order to extract coding knowledge. Machine Learning techniques manipulates features to produce an intelligent model (Collobert, Weston, 2008), consequently the problem is to determine which features could be extracted from the data to perform efficient learning. In the medical area, researchers extract feature matrices from medical reports and other conventional medical sources from patient episodes. Next, machine learning methods are applied on these matrices in order to generate models that can predict a diagnosis code. Different algorithms tackle this prediction problem: Decision Trees (Farkas, Szarvas, 2008), K-Nearest Neighbors (KNN) (Aronson *et al.*, 2007; Ruch *et al.*, 2007; Erraguntla *et al.*, 2012), Naïve Bayes Classifiers (Pakhomov *et al.*, 2006; Okamoto *et al.*, 2012), Regression (Xu *et al.*, 2007; Lita *et al.*, 2008), Support Vector Machine (SVM) (Yan *et al.*, 2010). Some mapping techniques are proposing to encode the disease through linking Medical dictionaries with international disease codes such as (Pereira *et al.*, 2006).

Some techniques use expert rules to achieve a high quality encoding. Researchers are transforming experts' coding knowledge into rules directly applied on the medical reports. An example is proposed by (Goldstein *et al.*, 2007), using hand crafted rules applied on radiology reports. The rules aim to extract lexical elements from radiology reports written in free text, lexical elements can be generated using semantic features to include negations, synonyms and uncertainty. The results of such techniques can reach interesting quality measures (for instance 88% F1 measure score) (Farkas, Szarvas, 2008). The limit of the methods applied on conventional data is that they are difficult to generalize in most of the cases.

(2) Structured data

Few works in the literature used structured patient data for diagnosis prediction. In such cases, data are mostly extracted from medical records, such as patient information (age, sex, length of stay), clinical information (prescription, medications) and other related medical data such as medical procedures and diagnoses. The interesting study of (Lecornu *et al.*, 2009) is based on statistical methods and probabilities. The authors focus on three types of medical data in order to estimate the probability of a diagnosis code. The first type is patient information (age, sex, length of stay), the second type is medical unit information and the third type is medical procedures. According to their study, diagnosis prediction is considered valid if it falls within the first 10 diagnoses ordered by probability score. The results of (Lecornu *et al.*, 2009) show that medical procedures were the most informative input, whereas the patient information was the least informative input. The authors report that better results could be achieved using all the inputs together by defining the right coefficient for each input. The limit of probabilistic and statistical approaches is the sensibility of these methods with respect to the quality of the used data. In particular, these methods generate imperfect results when they are applied on imperfect data, missing data or erroneous codes. Data mining approaches are good alternative, since data preprocessing techniques can help reducing the impact of imperfect data (Han *et al.*, 2012).

The authors of (Ferrao *et al.*, 2013) propose to use well structured data extracted from electronic medical records and convert them to around 5000 features. They use different data mining algorithms in several steps: naïve bayes and decision trees algorithms in (Ferrao *et al.*, 2012), SVM in (Ferrao *et al.*, 2013) and finally regression algorithms in (Ferrao *et al.*, 2015), trying to assign codes during different periods of the inpatient episode. All the proposed algorithms gave about similar evaluation of F1-measure but the results are still less effective than the F1-measure results reached by NLP techniques on radiology reports (Farkas, Szarvas, 2008; Goldstein *et al.*, 2007).

In France, two studies used data mining techniques to tackle the problem of assigning medical codes to inpatient episodes (Djennaoui *et al.*, 2015; Pinaire *et al.*, 2015). These approaches used other diagnoses occurred in previous inpatient episodes and constructed sequential patterns rules to predict a diagnosis code in the current patient episode. Two out of three diagnoses were successfully predicted using sequential patterns in (Djennaoui *et al.*, 2015).

For the data mining algorithm, we use decision trees, the main reason behind this choice is the interpretability of the model. The extracted model can be easily verified by domain experts such as physicians (Tuffery, 2007). In terms of performance, decision trees can produce a good prediction model using a similar data structure described in (Soni *et al.*, 2011). Although, other data mining algorithms might produce better models, we chose interpretability over performance. Moreover, decision trees are less sensitive to imbalanced datasets i.e. when the dataset contains unequally distributed classes (Cieslak, Chawla, 2008). Furthermore, the scalability of some versions of decision trees (Chrysos *et al.*, 2013) is an additional argument since we plan to experiment our methodology on the national dataset.

To tackle the problem of assigning secondary diagnoses codes to patient episodes, we aim to propose a general method that uses structured input and that avoids the ambiguities raised by conventional data. We are going to use the PMSI database. All the available structured variables in this medical database split into the categories mentioned by (Lecornu *et al.*, 2009; Ferrao *et al.*, 2012). In a previous work

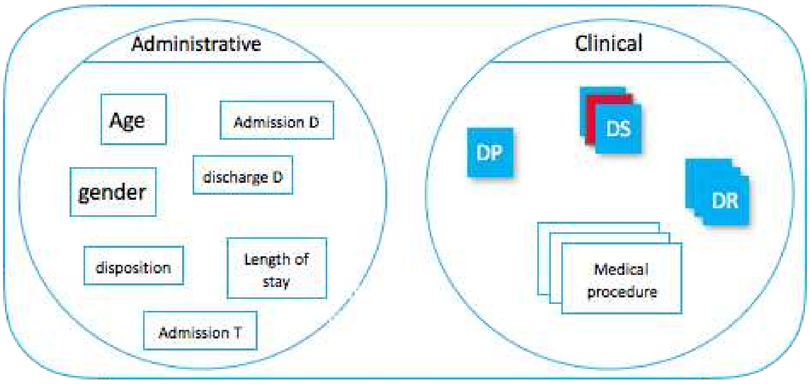

Figure 1. PMSI information type

(Chahbandarian *et al*., 2016), we used these data to highlight relevant information for the secondary diagnoses task, using decision trees and using goodness split metric to calculate the best features. In this work, we explore different representations of the medical data and different organisations of the subset of dataset used in learning process to increase the efficiency of the decision tree.

3. Preliminary

3.1. Materials

Data were obtained from a structured database: the PMSI. The PMSI database is described according to a common structure used at the regional scale as well as the national one. Our research is made in collaboration with the physicians in charge of the PMSI database of the "*Centre Hospitalier Intercommunal de Castres Mazamet*", a regional hospital in France. The first step of our research aims at processing data mining on this local database, the second step consists in a global validation on a regional PMSI database.

The PMSI contains anonymous discharge summaries (*Résumé de Sortie Anonymisé (RSA)*). Each summary characterises an inpatient episode through two types of information shown in Figure 1:

– *Administrative information* such as admission date, Discharge date, Admission mode, Discharge mode (transfer, death), Length of stay, Gender, Age. This category aggregates the patient information and medical unit information types described in (Lecornu *et al*., 2009).

– *Clinical information* such as the primary diagnosis (DP) that motivates the inpatient episode, secondary diagnoses (DS) on which we want to focus and related diagnoses (DR). It also contains all the medical procedures performed during the inpatient episode.

To encode diagnoses, the medical staff usually uses the *International Classification of Disease*[1] (ICD-10). In France, the medical staff uses the French version of ICD-10 named CIM-10[2]. This is a hierarchical classification: the first levels of organisation consist in chapters gathering same characteristic diseases (such as chapter II dedicated to tumoral diseases), categories help refining this classification. Currently, about 2,049 categories are commonly used for coding. The last level precisely describes each disease and the CIM-10 contains 33,816 codes in which the first three characters stand for code categories. To encode the medical procedures, the Common Classification of Medical Procedures CCAM (*Classification Commune des Actes Médicaux*)[3] is used. It is also a hierarchical classification: the first levels are made of the chapters which organise the medical procedures according to the medical system impacted (such as nervous system). There are around 1,700 standard medical procedure codes classified under 19 chapters.

Three kinds of information will be clearly distinguished in this paper:

– *Primary diagnoses (DP)* are the main diagnoses and are supposed to be rather easily encoded (through ICD-10) by healthcare professionals.

– *Secondary diagnoses (DS)* are sometimes difficult to detect but they are important to notice in order to get an exhaustive information on the performed care. They are also encoded in ICD-10.

– *Other features* such as administrative or clinical information are contained in PMSI database including the medical procedures encoded in CCAM.

3.2. Objectives

Medical databases are rich in data but poor in knowledge. Data mining is a way to extract previously unknown hidden data that could be useful. Machine Learning (ML) approaches provide tools and techniques that enable discovering knowledge from raw data. The idea is to identify relevant patterns in a database and to generate a model that can predict similar cases (Witten, Frank, 2005).

Our goal is to assess the suitability of data mining techniques for detecting a secondary diagnosis (DS), knowing the primary diagnosis (DP) and the other features encoded in PMSI (age, entry mode, medical procedures performed, etc.).

To alleviate the risk of encountering noisy data in the whole PMSI database, we chose to work on a subset of data where the DP is fixed. It is realistic since the DP is supposed to be easily known in our medical situation. Hence, the problem can be defined as follows:

In the subset of data, are we able to predict DS, knowing the other features?

[1] http://www.who.int/classifications/icd/

[2] http://www.atih.sante.fr/mco/presentation

[3] http://www.atih.sante.fr/version-39-de-la-ccam

The different databases were selected in three steps:

1. **Selection of interesting DS**: the physician in charge of the Medical Information Department (DIM) in the "*Centre Hospitalier Intercommunal de Castres Mazamet*" hospital proposed to focus on interesting and frequent secondary diagnoses which are difficult to detect as they are usually not well described across the medical sources. Seven DS were retained as listed in Table 1.

2. **Selection of frequent associated DP**: for each selected DS, the idea is to retain some frequent DP, where the DS is found as associated with the DP.

3. **Building the dataset**: the sub dataset, containing all the cases with the selected DP (associated or not with the DS) will constitute our learning database to predict DS, based on other features, once DP is encoded.

Table 1. The studied secondary diagnoses

ICD-10 codes	Labels
J96	Respiratory failure
B96	Bacterial agents such as Mycoplasma and pneumoniae
T81	Complications of procedures
R29	Nervous and musculoskeletal systems such as (Neonatal tetany)
R26	Abnormalities of gait and mobility
E66	Overweight and obesity
E44	Malnutrition

4. Methodology

The choice of ML technique and the selection of data used play a fundamental role in the quality of the results of the proposed application. ML techniques build the model based on some input data called features. This model makes prediction outputs called labels. In order to produce better prediction outputs, the features should be chosen carefully. This means that when preparing feature data, we should handle missing data, discretize the continuous numeric values and represent features.

The first step of our work consists of a feature selection step where the important features are identified regardless of the studied diagnoses, the second step queries the sub-dataset that corresponds to the studied DP-DS and provides personalised data preprocessing to this sub-dataset, namely discretizing continuous features and sampling the imbalanced datasets. The next step uses the prepared sub-dataset to build a prediction model, in our case a decision tree. The final step evaluates the prediction model. In the following sections more details are provided for each step.

4.1. Feature representation

For our problem, all the features are extracted from PMSI database. ML techniques needs to choose the appropriate feature representation for the ML model. We can categorise the features into three sets:

– The first set of features is composed of **personal information** which includes the patient's gender and his/her age at admission. We discarded the zip code, as all the patients come from the same area where the hospital is located, but this information would be interesting to investigate in case of using the national version of PMSI database.

– The second set of features concerned **inpatient episode** including the length of stay, the patient admission type, the patient discharge status, the time interval between the admission date and the date of the first medical procedure, the transfer count between medical units during the inpatient stay, the medical procedures count, the season of the admission and the previous inpatient episode count calculated thanks to a process of anonymous chaining available in the PMSI databases which permits to link information from a single patient.

– The third set of features is **derived from medical information**, more particularly from inpatient diagnoses and medical procedures. Each feature in this set represents the presence or absence of medical information in the inpatient episodes. Diagnoses and medical procedures can be classified using different levels of granularity or hierarchy and using different levels of classification can limit the number of derived features. As (Sebban *et al.*, 2000) says that extra large number of features does not yield necessarily to good results for ML learning algorithms especially for decision trees case, we want to explore the quality and the speed of the ML algorithm when using different number of features. Therefore, we can use the hierarchical representation of medical information to focus on different granularity level of representation in order to control the number of features used in the model.

- Concerning medical procedure features, we have derived 19 features, each feature indicating whether one or many medical procedures in the corresponding chapter have occurred during the inpatient episode.

- Concerning diagnoses features, we have derived 145 features according to two levels of granularity. **(1) Coarse level** granularity which contains 19 chapters of diagnoses classification and **(2) Fine level** granularity which contains 126 specific chapters of diagnoses classification.

– Finally, the output label of the ML model is a boolean output, positive if the ICD-10 code exists in the inpatient episode and negative otherwise.

In total, we have used 181 features to build our ML model. A detailed description can be found in Table 2.

Table 2. Used features in the decision trees

	Variables		
	Name	**Description**	**Valid values**
Personnal	Gender	Patient’s gender	F=Female, M=Male
	Age	Patient’s age at admision	Below; Mean; Overs
Inpatient variables	Length of stay	Time interval between admission date and discharge date	Below; Mean; Over
	Admission type	Patient’s admission type	1=Emergency 2=Urgent 3=Elective 4=Newborn 5=Trauma 9=Information not available
	Disposition	Patient’s discharge status	1= Discharge to home 2=Transferred to short-term facility 3=Transferred to skilled nursing facility 4=Transferred to intermediate 5=Transferred to other health care facility 6=Transferred to home health care 7= Left AMA (Against Medical Advice) 20= Expired/Mortality
	Season	The season at the admission	Summer Winter Fall Spring
	Frequency	The count of the inpatient episodes of the patient during his life	Below; Mean; Over
	Delay	Time interval between admission date and first medical procedure	Below; Mean; Over
	Inpatient transfer count	The count of the transfers between medical units in the inpatient episode	Below; Mean; Over

	Variables		
	Name	**Description**	**Valid values**
	Medical procedures count	The count of the medical procedure during the inpatient episode	Below; Mean; Over
Derived flags	Classified	A flag indicating whether the inpatient stay has a classified/ important medical procedure or not.	0=No 1=Yes
	Emergency	A flag indicating whether the inpatient stay has an emergency case or not.	0=No 1=Yes
	Medical procedure groupings	19 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding medical procedure category.	0=No 1=Yes
	Urgent medical procedure grouping	5 flags, each flag indicates whether the inpatient stay has a medical procedure within the corresponding urgent medical procedure category.	0=No 1=Yes
	Coarse level diagnoses granularity	19 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding diagnosis granularity.	0=No 1=Yes
	Fine level diagnoses granularity	126 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding diagnosis granularity.	0=No 1=Yes
Output	Label	A flag indicating whether the inpatient stay has the studied secondary diagnosis or not.	0=Negative 1=Positive

4.2. Data preprocessing

Data preprocessing is an important step in the data mining process. It is particularly important because low quality data could lead to low quality mining results (Han *et al.*, 2012).

In this work, data preprocessing should address two issues:

– The first issue is dealing with the continuous numerical data in the PMSI database.

– The second one is dealing with the imbalanced repartition of the positive and negative examples.

Missing data is not addressed in this work: all the features we used are required for every inpatient stay in PMSI furthermore we have verified that there are not any missing data in our dataset.

Continuous numerical data: In PMSI database, there are two kinds of data: numerical (continuous or discrete) and categorical. Decision trees are efficient with categorical values. Consequently, if the values are numeric then they are discretized prior to build the model (Tuffery, 2007). In the literature several methods are proposed to discretize numeric values into categorical. For instance, "binning" is an unsupervised method which discretizes the numerical values either into equal-interval binnings or into equal-frequency binnings. Supervised discretization methods, such as "entropy-based", measure the information gain to the class and split the intervals recursively (Witten, Frank, 2005). Although these methods could generate a model with good performance, the intervals used to build it lack clarity in terms of interpretability in a first test we performed without treating this problem, ages for example were splitted into the following intervals (>6),([7-12]),([13-30]),([31-40]),(<40) such intervals could not really make sense for medical interpretation. It is important in the medical domain to help the physicians to interpret the results. Assuming that all the continuous features are normally distributed, we have chosen to discretise the continuous features into three intervals ("**Below**", "**Mean**", "**Over**"), these intervals change according to each couple (DS-DP) of secondary and primary diagnoses studied.

– "**Below**" refers to values smaller than the mean minus one standard deviation.

– "**Mean**" refers to data between the mean plus minus one standard deviation.

– "**Over**" refers to data above mean plus one standard deviation.

The features, which have been discretized, are: frequency, transfer count, medical procedure count, diagnoses count, age, length of stay and delay.

Imbalanced dataset: "A dataset is imbalanced if the classification categories are not approximately equally represented" (Chawla, 2005). Real life datasets are often imbalanced, this is particularly true in the medical databases, where certain studied diagnosis tend to be the minority class (Rahman, Davis, 2013). "A well balanced dataset is very important for creating a good prediction model" (Rahman, Davis, 2013). For instance, in case of respiratory failure secondary diagnosis study, our dataset contains

4166 records with this diagnosis (positive outputs) out of 90,000 inpatient stays, the ratio of positive outputs is then 5%.

There are three main sampling methods in the literature to tackle the problem of imbalanced dataset:

– **Undersampling**: it removes samples from the majority class using an undersampling algorithm, such as Tomek Links (Tomek, 1976), Condensed Nearest Neighbor Rule (Angiulli, 2005) or the baseline method Random under-sampling.

– **Oversampling**: it generates new samples from the minority class using an oversampling algorithm, such as SMOTE (Chawla *et al.*, 2002) or the baseline method Random over-sampling.

– **Cost-Sensitive**: it takes the misclassification cost into consideration, such as MetaCost (Domingos, 1999), Costing (Zadrozny *et al.*, 2003) or the baseline method weighting (Ting, 1998) where the minority and majority are given classes different weights in order to optimise the misclassification cost (Elkan, 2001).

In our work we chose the baseline methods of undersampling and cost-sensitive, namely Random under-sampling and weighting methods, since they are effective and they don't cost much calculation power compared to oversampling methods since they tend to add more data that need to be processed. Moreover, oversampling methods could add bias in the medical data and tend to perform worse than undersampling methods (Drummond, Holte, 2003).

4.3. Decision tree

Among the machine learning methods, we have chosen to use decision tree. This method belongs to the class-labeled training tuples. We chose the decision tree method because it generates simple models, it is easy to interpret and it can be validated by physicians who are not necessarily experts in computer science. Furthermore, decision tree method is scalable and can produce efficient models even when large amounts of data are used. Finally, decision trees are less sensitive to imbalanced datasets (Cieslak, Chawla, 2008). Decision trees use an attribute selection rule at each node of the tree to split the data (split criterion), this rule is important to classify the records correctly. The main split criteria in the literature are Information Gain and Gini Index (Han *et al.*, 2012). The difference in the performance between those two criteria is not huge. The best criterion is debatable and it depends on the used dataset (Raileanu, Stoffel, 2004). Since Gini Index tends to perform a bit faster than Information Gain (Raileanu, Stoffel, 2004), we retained Gini Index. For the decision tree, we have chosen the *Classification and Regression Tree* (CART) algorithms (Breiman *et al.*, 1984) that uses Gini Index. CART is a binary decision tree, which is built by recursively splitting each node into two child nodes, until there is no significant decrease in the Gini Index criterion.

Overfitting problem occurs when the model is more accurate on the training set than on the testing data. Pruning can be used to avoid the overfitting problem (Han *et al.*, 2012). The minimal cost-complexity pruning is implemented in the CART decision tree as described in (Breiman *et al.*, 1984). Default parameters for pruning were used in our case because such overfitting problem could occur.

CART decision tree was then used to perform our objective: to know if we could help to detect a secondary diagnosis (DS), knowing the primary diagnosis (DP) and the other features encoded in PMSI (age, entry mode, medical procedures performed, ...). In order to evaluate which decision tree could be the best, we compared the performances of different decision trees, each one being built using different choices according to following points:

– Granularity level: as the codification of the diagnoses belongs to a hierarchical classification, it is possible to use different levels of description: either coarse level with 19 features (which correspond to general chapters) or fine level of diagnoses with 126 features (more specific chapters).

– imbalanced dataset: as the PMSI database contains by nature more negative examples than positive ones, we have made the hypothesis that the we can build a better performance decision tree by balancing the number of positive and negative examples. To verify the hypothesis we consider three sampling methods.

- The first method uses the original dataset without any sampling method.
- The second method gives the positive examples in the dataset double weight compared to the negative ones.
- The third method uses randomly undersampling technique with 1:1 ratio.

Few preliminary test helped us to choose the weightings and the ratio of random undersampling presented in this work we still have to improve these choices: after determining the best sampling method we plan to try different tunings and to do excessive empirical study in order to determine the best choices.

This choices of granularity level and of sampling methods were tested through different situations, as presented in Table 3 which classifies the tested use cases.

Table 3. The tested use cases - Each situation was named as test x and will be described in experiments section

	Dataset		
	Original	Cost sensitive	Random sampling
Fine level of granularity	Test1	Test3	Test5
Coarse level of granularity	Test2	Test4	Test6

4.4. Algorithm

The general algorithm followed to build and to evaluate the decision tree used to predict secondary diagnoses is described in Algorithm 1.

The first step (1->3) allows to choose the right configuration by fixing 3 variables:

– The weight of positive and negative examples (for instance, we decide to weight a positive example twice in order to highlight its importance).

– The random undersampling option.

– The granularity level of diagnosis (for instance, we choose a decision tree based on the 19 features issued from general chapters).

The second step (4) queries the most 10 frequent Primary Diagnoses DPs occurred with the studied secondary diagnosis. (for example, in case of "B96" bacterial agents infection as DS, the most frequent primary diagnoses found in the database are "**Acute tubulointerstitial nephritis**" with the code "N10", "**Malaise and fatigue**" with the code "**R53**", "**Fever**" with the code "**R50**", etc...) Afterwards, (6) for each DP we query the corresponding dataset that contains the positive and negative examples. Then, we do all the preprocessing (7->9), next (10) we split the data into K training and testing sets and for each set (12) we use the training set to build a decision tree. Afterwards (13), the tree is pruned in case the subtree produces better performance. We evaluate (15) the tree using the testing set. Next (16), we average the evaluations produced by each fold. Finally (18), we average the evaluations of all the performances of the decision trees of the Primary Diagnoses DPs.

Algorithm 1 The steps followed to build secondary diagnoses decision tree

```
 1: Set(The weight of positive and negative examples)
 2: Set (Undersampling option)
 3: Set (Granularity level of diagnoses)
 4: Query the most 10 frequent DPs occurred with the DS
 5: for each primary diagnosis DP do
 6:     Query the dataset using the chosen granularity level
 7:     Discretize the continuous features (age-length of stay - frequency - medical
        procedures count...)
 8:     If undersampling selected randomly undersample the majority class
 9:     Give the positive and negative classes their weights
10:     Split the data into k folds
11:     for Each fold do
12:         Build the decision tree with the training set using CART algorithm
13:         Prune the tree (if possible)
14:         Evaluate the model by Measuring (Precision -Recall- F1) on the testing set
15:     end for
16:     Calculate the average evaluations of the folds
17: end for
18: Calculate the average of the evaluations of DPs
```

4.5. Evaluation

The results were evaluated using 5-fold cross validation as the dataset does not contain sufficient positive records to dedicate independent testing set. In each fold, we divided the dataset into 80% training set and 20% testing set. We used the standard metrics used to evaluate classification Precision, Recall and F1-measure.

The measurements are defined based on the following sets according to (Tuffery, 2007):

– **TP** is the number of True Positive instances, which represent instances that are correctly assigned to positive examples.

– **TN** is the number of True Negative instances, which represent instances that are correctly assigned to negative examples,

– **FP** is the number of False Positive instances, which represent instances that are incorrectly assigned to positive examples,

– **FN** is the number of False Negative instances, which represent instances that are incorrectly assigned to negative examples.

Precision is the ratio of correctly assigned examples to the total number of examples produced by the classifier.

$$P = \frac{TP}{(TP + FP)} \tag{1}$$

Recall is the ratio of correctly assigned examples to the number of target examples in the test set.

$$R = \frac{TP}{(TP + FN)} \tag{2}$$

F1-measure represents the harmonic mean of precision and recall according to the formula in (3):

$$F1 = \frac{2P * R}{(P + R)} \tag{3}$$

Using these measurements we aimed to evaluate three aspects:

– The credibility of the automatic prediction, using decision tree method, of the DSs assignment codes in the inpatient episode.

– The impact of balancing datasets to answer question of limiting the effect of the large number of negative examples compared to the positive ones.

– The impact of the diagnoses granularity: does one of the levels produce a better performing decision tree (fine level when specific diagnoses groupings are considered, coarse level when general diagnoses groupings are considered).

5. Experiments

This section provides a detailed description of the dataset used and describes the implementation of the proposed approach, in addition to different scenarios to try different combination of the parameters to choose the best use case of the methodology described in the Table 3.

5.1. Dataset

In order to evaluate our method, we have used an anonymized sample data extracted from the PMSI database of "Centre Hospitalier Intercommunal de Castres Mazamet" hospital. It contains around 90,000 inpatient episodes between 2011 and 2014. In Table 4, we have detailed the number of cases for each studied diagnosis and we have presented the most 10 frequent Primary Diagnoses DP found as associated with this DS.

Table 4. The studied Secondary Diagnoses DSs and the most 10 frequent Primary Diagnoses DPs associated with the DS

ICD-10 codes	**Labels**	**Count in PMSI DB**	**Most Frequent DP ordered by frequency**
J96	Respiratory failure	4166	I50(Heart failure)-R06(Abnormalities of breathing)-J96(Respiratory failure)-J44(obstructive pulmonary disease)-J18(Pneumonia)-R53(Malaise and fatigue)-J20(Acute bronchitis)-J15(Bacterial pneumonia)-Z51(Encounter of medical care)-J69(Pneumonitis)
B96	Bacterial agents infection such as Mycoplasma and pneumoniae	6514	N10(Acute pyelonephritis)-R53(Malaise and fatigue)-I50(Heart failure)-R50(Fever)-R10(Abdominal and pelvic pain)-R06(Abnormalities of breathing)-J44(obstructive pulmonary disease)-J96(Respiratory failure)-N41(Inflammatory diseases of prostate)-J18(Pneumonia)
T81	Complications of procedures	1150	Z48(Encounter for attention to dressings, sutures and drains) -L02(Cutaneous abscess)-C18(Malignant neoplasm of colon) -S72(Fracture of femur) -K56(Paralytic ileus) -K65(Peritonitis) -K80(Cholelithiasis)-R10(Abdominal and pelvic pain)-K63(intestine disease) -Z43(Encounter for attention to tracheostomy)
R29	Nervous and musculoskeletal systems	1596	R53(Malaise and fatigue)-S06(Intracranial injury) -F05(Delirium) -R29(Nervous and musculoskeletal systems) -R41(Disorientation) -I50(Heart failure) -I95(Hypotension) -R52(acute) -S72(Fracture of femur) -I63(Cerebral infarction)
R26	Abnormalities of gait and mobility	2378	R53(Malaise and fatigue)-I50(Heart failure)-R06(Abnormalities of breathing)-J20(Acute bronchitis)-J69(Pneumonitis)-R50(Fever)-J15(Bacterial pneumonia)-F05(Delirium)-Z51(Encounter of medical care)-J18(Pneumonia)
E66	Overweight and obesity	5453	I50(Heart failure) -R07(Chest pain)-E11(Type 2 diabetes mellitus) -R06(Abnormalities of breathing) -J96(Respiratory failure) -R53(Malaise and fatigue) -R10(Abdominal and pelvic pain) -I48(Atrial fibrillation and flutter)-Z48(Encounter for attention to dressings, sutures and drains)-K80(Cholelithiasis)
E44	Malnutrition	2144	R53(Malaise and fatigue)-F05(Delirium)-I50(Heart failure)-R29(Nervous and musculoskeletal systems)-R06(Abnormalities of breathing)-J18(Pneumonia)-R41(Disorientation)-R10(Abdominal and pelvic pain)-J44(obstructive pulmonary disease)-J15(Bacterial pneumonia)

5.2. Implementation

The proposed algorithm is implemented using R-Studio[4] and weka[5]. R-Studio is used to query the subsets from MySql[6] database where the PMSI is stored, then the preprocessing of the dataset is performed using R-Studio, next a dataset with the ARFF[7] format is produced, An ARFF (Attribute-Relation File Format) file is a text file that contains features description in addition to the dataset instances in a special format mostly used with weka. Finally, weka platform is used to build the CART decision tree, as shown in Figure 2.

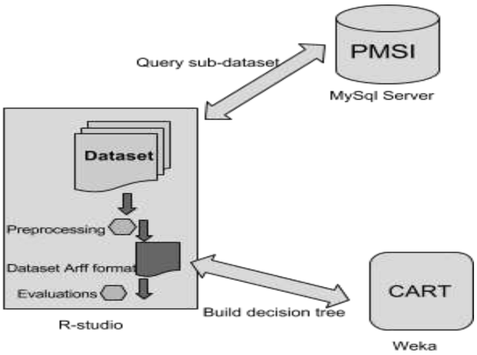

Figure 2. Implementation of the Algorithm

We have experimented our approach according to the use cases described in the Table 3 in three scenarios. In each scenario we represent features as described in the section 4.1 which consists of coarse and fine level of diagnoses granularity. Moreover, we have changed the methods for sampling imbalanced data set. Hence, the three scenarios can be described as following:

– **Scenario 1** corresponds to using the original dataset without any sampling. Test 1 & 2 in the Table 3.

– **Scenario 2** corresponds to cost-sensitive learning method for sampling imbalanced dataset. Test 3 & 4 in the Table 3.

– **Scenario 3** corresponds to random undersampling of negative examples to a ratio of (1:1). Test 5 & 6 in the Table 3.

Scenario 1. Figure 3 summarizes the results of the different measures on the original dataset. First, we can observe that even for fine and coarse granularity, using all the dataset is not an interesting strategy as recall and F1-measures results are very low. Except for B96 (bacterial agents) and J96 (Respiratory failure), our results approxi-

[4] https://www.rstudio.com/products/rstudio/

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] https://www.mysql.fr/

[7] https://weka.wikispaces.com/ARFF

mate 2%. For B96 and J96, we can observe that the results of fine granularity are more interesting than the results of coarse level granularity.

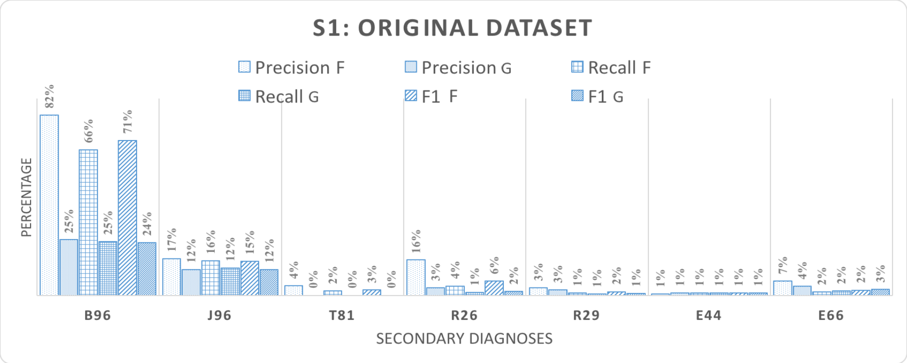

Figure 3. Summary of the average measurements of the decision tree's performance in the scenario 1 - based on original dataset - using fine and coarse levels of granularity for all the studied diagnoses, F: Fine Level; C: Coarse Level

Scenario 2. In the lights of the results of the evaluations shown in Figure 4, we can observe that the measurement varied between different diagnoses. On one hand, B96 scored the best F1, precision and recall measurements around 75%. On the other hand, other diagnoses scored lower percentages using the same measurements. As reported by Stanfill (Stanfill *et al.*, 2010), a same ML applied on different diagnoses, produces different results. Our results confirm such a variation of measurements, and the complexity of the problem. Concerning the highlighted issues about the effect of the granularity level, we notice that using fine level granularity gives better measurements compared to using coarse level granularity. We can observe that the differences between fine and coarse level of granularity range between 1% and 27% in the results Figure 4. In particular, for B96 we notice an important enhancement of results quality using the fine level granularity.

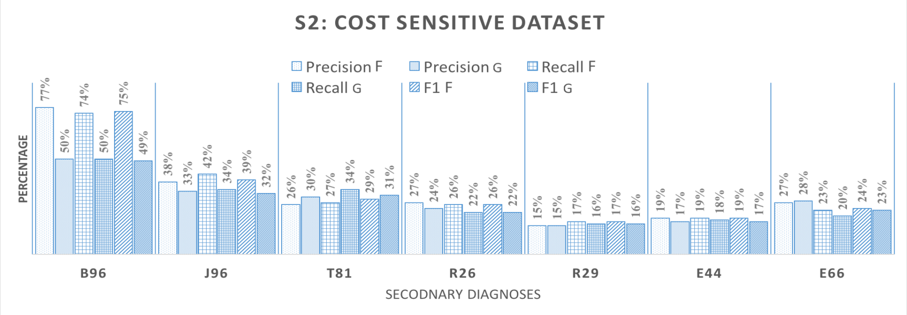

Figure 4. Summary of the average measurements of the decision tree's performance in the scenario 2 - based on cost sensitive learning - using fine and coarse levels of granularity for all the studied diagnoses, F: Fine Level; C: Coarse Level

Scenario 3. Figure 5 shows the results of the third scenario. Clear improvement can be observed in the quality of detection of all secondary diagnoses. Compared to the results presented in Figures 3 and 4 in which the used sampling methods privileged B96 and J96 diagnoses, this evaluation substantiates that sampling negative examples according to 1:1 ratio is the best method to predict a right quality over all type of secondary diagnoses. In fact, the results show that the values of the quality measures range between 55% and 84%, which are very trustworthy to satisfy our main objective. The difference of each sampling methods can be observed clearly in Figures 6, 7 and 8, each figure shows the performance and the differences between the sampling methods using the metrics F1, Precision and recall in order.

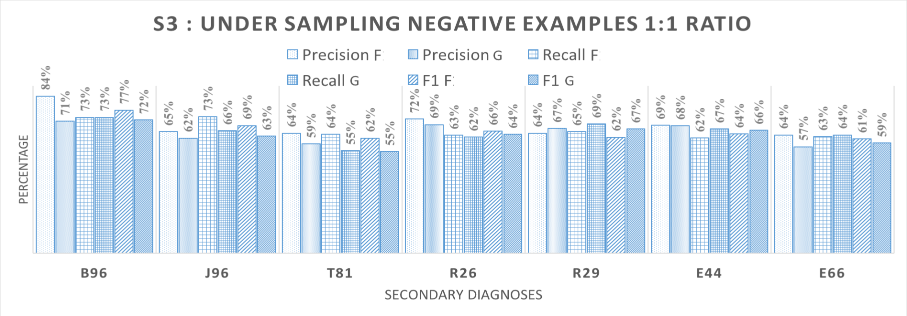

Figure 5. Summary of the average measurements of the decision tree's performance in the scenario 3 - based on undersampled dataset - using fine and coarse levels of granularity for all the studied diagnoses, F: Fine Level; C: Coarse Level

To sum up the differences between the performed experimentations, we overlap the results of the three scenarios on the three metrics F1-measure, recall and precision respectively in Figures 6, 7 and 8. The most important remarks are:

– Fine level granularity features give better results than coarse level granularity features regardless the type of secondary diagnoses and the type of metric, this seems coherent with the fact that detailed level provide more information and can give better prediction power.

– The method of sampling impacts the quality of results. We can observe that the under sampling method improves the results significantly compared to the cost sensitive and the original unsampled dataset regardless the type of secondary diagnoses and the type of metric. Intuitively, sampling methods are improving the quality because they make the number of positive examples more representative compared to negative examples.

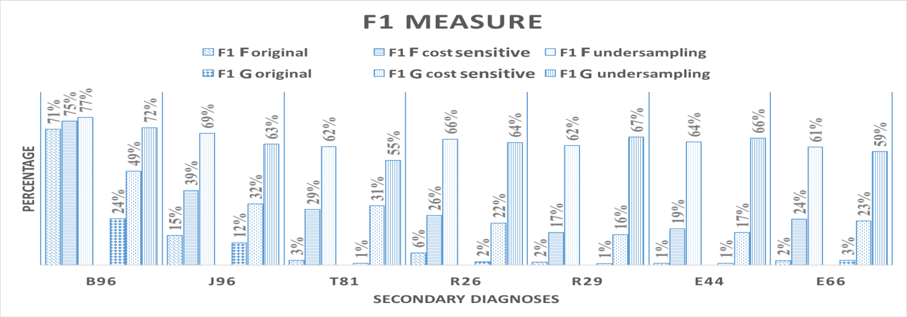

Figure 6. F1 measurement based on all the three sampling methods

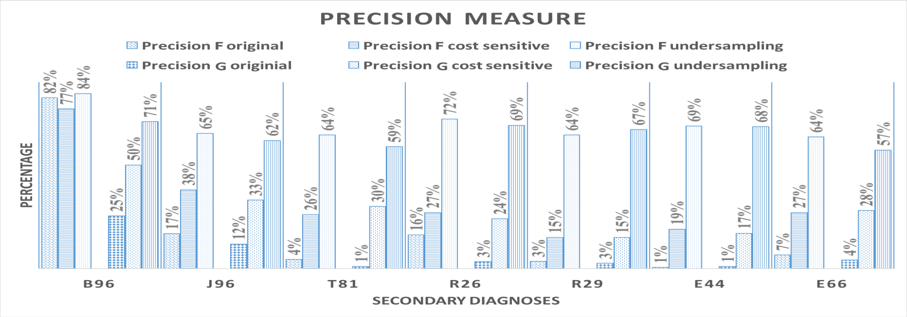

Figure 7. Precision measurement based on all the three sampling methods

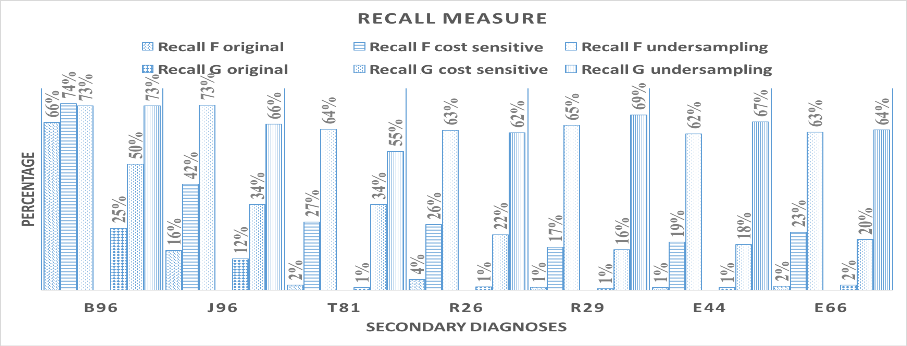

Figure 8. Recall measurement based on all the three sampling methods

6. Discussion

The main objective is to know if data mining techniques could help to detect a secondary diagnosis (DS), knowing the primary diagnosis (DP) and the other features encoded in PMSI (age, entry mode, medical procedures performed ...). Therefore, we selected an approach using decision trees that has promising result and can satisfy our main objective to help the coders and notify them whenever a secondary diagnosis is missing by counting on structured data only. Our hypothesis of fixing the primary diagnosis has helped to enhance the prediction performance.

We raised secondary issues to verify their impact on the decision tree. Concerning the granularity level, as the codification of the diagnoses belongs to a hierarchical classification, it is possible to use different levels of description: either coarse level with 19 features (which correspond to general chapters) or fine level of diagnoses with 126 features (more specific chapters). We compared the performances of two decision trees, each one is built using different level of diagnoses granularity. The results showed that by using the fine level of granularity we can enhance on average 5% to 10% all the quality measures regardless of the predicted diagnosis code. The prediction power seems to be related to the preciseness of the medical information.

Some diagnoses had better performance decision tree compared to others such as B96 "bacterial agents". B96 is the most frequent secondary diagnosis. The performance could be explained either by the fact that the ratio between positive and negative examples is the best one in our database, or by medical specificity of bacterial agents. A better understanding of predictive power of each feature could be established with the help of the medical staff in the hospital. The understanding of the feature could explain the behavior and the performances of each model.

Concerning the imbalanced dataset, as the PMSI database contains by nature more negative examples than positive ones, the improvement of results in the third scenario when the balanced dataset is used confirms that balancing techniques are useful to produce better performance decision trees. In the second scenario cost sensitive learning is used by giving the positive examples in the dataset double weight compared to the negative ones, this technique produced 25% better performance model compared to the model based on original dataset. Finally, we used random undersampling technique to reduce the number of negative examples to be equal with the positive ones, this technique generated the best performance model regardless to the predicted diagnosis which can be a good step towards better application.

The strength of the approach is to provide a generic structured dataset that can be populated with any PMSI database, while allowing personalized data preprocessing for each studied (DP-DS). Such approach includes customized discretization ranges for the continuous features adapted to each subset of data, to provide better interpretability and to improve the prediction quality. The weakness of the approach is that in some cases the queried subset is insufficient to build a model. This could be solved by using larger databases, such as the national version of PMSI. Another weakness is that our approach is not optimised in terms of memory consumption and training time, due to the large number of features. We hope to solve that in the future work by including automatic features selection methods, such as correlation feature selection method CFS (Hall, 1998), which helps to choose only the relevant set of features.

7. Conclusion

The paper outlined preliminary results of our methodology to develop an automatic model able to assign secondary medical codes. The proposed approach consists of a model based on decision tree, which is built on structured data extracted from the PMSI database. The performance of the model ranged from 61% to 77% F1 measure. Therefore, the proposed methodology holds great promise for improving the tedious task of coding secondary diagnoses.

We experimented our approach according to three scenarios to address the imbalanced dataset problem. In the first scenario the original dataset was used without any sampling, in the second scenario cost sensitive learning was applied, and in the third scenario random undersampling of negative examples was applied. In each scenario we used different diagnoses representation coarse and fine level of diagnoses granularity. The best performance model was achieved by using the third scenario choosing the fine level granularity of diagnoses representation.

For future perspectives, we plan to test the effect of using different level of medical procedures representations. Thanks to the common structure of PMSI database at the regional and national scales, we plan to extend our methodology and evaluate it on different scales of PMSI database. Moreover, we plan to explore new methods in order to balance the positive and negative examples in the training set as well as automatic methods for feature selection. Furthermore, new evaluation methods will be tested, taking into consideration imbalanced databases such as (Weng, Poon, 2008). Finally, we plan to evaluate our work in real world application and have a feedback from the users of our proposition.

Acknowledgements

The Ph.D. work of one author is funded by a grant from Midi-Pyrénées region, Castres-Mazamet Technopole, INU Champollion and Paul Sabatier University.

References

Angiulli F. (2005). Fast Condensed Nearest Neighbor Rule. In *Proceedings of the 22nd international conference on machine learning*, pp. 25–32.

Aronson A. R., Bodenreider O., Demner-Fushman D., Fung K. W., Lee V. K., Mork J. G. *et al.* (2007). From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. , pp. 105–112.

Breiman L., Friedman J., Olshen R. A., , Stone C. (1984). *Classification and Regression Trees*. Taylor & Francis.

Busse R., Geissler A., Quentin W. (2011). *Diagnosis-Related Groups in Europe: Moving towards transparency, efficiency and quality in hospitals*. McGraw-Hill Education (UK).

Chahbandarian G., Souf N., Bastide R., Steinbach J.-C. (2016). Increasing Alertness while Coding Secondary Diagnostics in the Medical Record. In *Proceedings of the 9th international joint conference on biomedical engineering systems and technologies*, pp. 490–495.

Chawla N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data mining and knowledge discovery handbook*, pp. 853–867. Springer.

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

Chrysos G., Dagritzikos P., Papaefstathiou I., Dollas A. (2013, jan). HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system. *ACM Transactions on Architecture and Code Optimization*, Vol. 9, No. 4, pp. 1–25.

Cieslak D. A., Chawla N. V. (2008). Learning Decision Trees for Unbalanced Data. In *Machine learning and knowledge discovery in databases*, Vol. 5211 LNAI, pp. 241–256. Berlin, Heidelberg, Springer Berlin Heidelberg.

Collobert R., Weston J. (2008, jul). A unified architecture for natural language processing. In *Proceedings of the 25th international conference on machine learning - icml '08*, pp. 160–167. New York, New York, USA, ACM Press.

Djennaoui M., Ficheur G., Beuscart R., Chazard E. (2015). Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes. *Studies in health technology and informatics*, Vol. 210, pp. 419–23.

Domingos P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining*, pp. 155–164.

Drummond C., Holte R. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop on Learning from Imbalanced Datasets II*, pp. 1–8.

Elkan C. (2001). The foundations of cost-sensitive learning. In *Ijcai international joint conference on artificial intelligence*, Vol. 17, pp. 973–978.

Erraguntla M., Gopal B., Ramachandran S., Mayer R. (2012). Inference of Missing ICD 9 Codes Using Text Mining and Nearest Neighbor Techniques. In *2012 45th hawaii international conference on system sciences*, pp. 1060–1069. IEEE.

Farkas R., Szarvas G. (2008, jan). Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, Vol. 9 Suppl 3, pp. S10.

Ferrao J. C., Healthcare S., Janela F. (2015). Predicting length of stay and assignment of diagnosis codes during hospital inpatient episodes. In *Proceedings of the first karlsruhe service summit workshop-advances in service research, karlsruhe, germany, february 2015*, Vol. 7692, p. 65.

Ferrao J. C., Janela F., Oliveira M., Martins H. (2013, sep). Using Structured EHR Data and SVM to Support ICD-9-CM Coding. In *2013 ieee international conference on healthcare informatics*, pp. 511–516. IEEE.

Ferrao J. C., Oliveira M. D., Janela F., Martins H. M. G. (2012, oct). Clinical coding support based on structured data stored in electronic health records. In *2012 ieee international conference on bioinformatics and biomedicine workshops*, pp. 790–797. IEEE.

Fetter R. B. (1991). Diagnosis Related Groups: Understanding Hospital Performance. *Interfaces*, Vol. 21, No. 1, pp. 6–26.

Goldstein I., Arzrumtsyan A., Uzuner O. (2007). Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 279–83.

Hall M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Unpublished doctoral dissertation, University of Waikato, Hamilton, New Zealand.

Han J., Kamber M., Pei J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.

J.Cios K., Moore G. (2002). Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine Journal*, Vol. 26, No. 1, pp. 1–24.

Lecornu L., Thillay G., Le Guillou C., Garreau P. J., Saliou P., Jantzem H. *et al.* (2009). REFEROCOD: a probabilistic method to medical coding support. In *Engineering in medicine and biology society, 2009. embc 2009. annual international conference of the ieee*, pp. 3421–3424.

Lita L. V., Yu S., Niculescu S., Bi J. (2008). Large Scale Diagnostic Code Classification for Medical Patient Records. In *Proceeding sof the international joint conference on natural language processing (ijcnlp'08)*, pp. 877–882.

Okamoto K., Uchiyama T., Takemura T., Adachi T., Kume N., Kuroda T. *et al.* (2012). Automatic Selection of Diagnosis Procedure Combination Codes Based on Partial Treatment Data Relative to the Number of Hospitalization Days. *Proc. APAMI 2012*, No. 4, pp. 1031.

Pakhomov S. V. S., Buntrock J. D., Chute C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association : JAMIA*, Vol. 13, No. 5, pp. 516–25.

Pereira S., Névéol A., Massari P., Joubert M., Darmoni S. (2006). Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. In *Studies in health technology and informatics*, Vol. 124, pp. 845–50.

Pinaire J., Rabatel J., Azé J., Bringay S., Landais P. (2015). *Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde.*

Potignon C., Musat A., Hillon P., Rat P., Osmak L., Rigaud D. *et al.* (2010). P146-Impact financier pour les établissements hospitaliers du mauvais codage PMSI de la dénutrition et de l'obésité. Étude au sein du pôle des pathologies digestives, endocriniennes et métaboliques du CHU de Dijon.

Rahman M. M., Davis D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, Vol. 3, No. 2, pp. 224–228.

Raileanu L. E., Stoffel K. (2004). Theoretical comparison between the Gini Index and Information Gain criteria. *Annals of Mathematics and Artificial Intelligence*, Vol. 41, No. 1, pp. 77–93.

Ruch P., Gobeill J., Tbahriti I., Tahintzi P., Lovis C., Geissbühler A. *et al.* (2007). From clinical narratives to ICD codes: automatic text categorization for medico-economic encoding. *Swiss Medical Informatics*, Vol. 23, No. 61, pp. 29–32.

Sebban M., Nock R., Chauchat J. H., Rakotomalala R. (2000). Impact of learning set quality and size on decision tree performances. *International Journal of Computers, Systems and Signals*, Vol. 1, No. 1, pp. 85–105.

Soni J., Ansari U., Sharma D., Soni S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43–48.

Stanfill M. H., Williams M., Fenton S. H., Jenders R. A., Hersh W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA*, Vol. 17, No. 6, pp. 646–51.

Ting K. M. (1998). Inducing cost-sensitive trees via instance weighting. In *European symposium on principles of data mining and knowledge discovery*, pp. 139–147.

Tomek I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics 6*, Vol. 6, pp. 769–772.

Tuffery S. (2007). Data mining et statistique décisionnelle : l'intelligence des données.

Weng C. G., Poon J. (2008). A new evaluation measure for imbalanced datasets. In *Conferences in research and practice in information technology series*, Vol. 87, pp. 27–32. Darlinghurst, Australia, Australia, Australian Computer Society, Inc.

Witten I. H., Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xu J. W., Yu S., Bi J., Lita L. V., Niculescu R. S., Rao R. B. (2007). Automatic medical coding of patient records via weighted ridge regression. In *Proceedings - 6th international conference on machine learning and applications, icmla 2007*, pp. 260–265.

Yan Y., Fung G., Dy J. G., Rosales R. (2010). Medical Coding Classification by Leveraging Inter-Code Relationships. In *Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining (kdd)*, pp. 193–201.

Zadrozny B., Langford J., Abe N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Data mining, 2003. icdm 2003. third ieee international conference on*, pp. 435–442.