جامعة أبو بكر بلقايد- تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

pour l'obtention du Diplôme de

MASTER en GENIE BIOMEDICAL

Spécialité : **Informatique Biomédicale**

présenté par : BOUKHOBZA Khadidja

A variable importance measure for a Cost Sensitive Random Forest Predition on a Budget

Soutenu le 26 juin 2019 devant le Jury

M.	CHIKH Mohammed Amine	*Prof*	Université de Tlemcen	Président
Mme	SETTOUTI Nesma	*MCA*	Université de Tlemcen	Encadreur
Mme	SAIDI Meryem	*MCB*	ESM de Tlemcen	Co-encadreur
M.	EL HABIB DAHO Mostafa	*MCB*	Université de Tlemcen	Examinateur
Melle.	GUILAL Rima	*Doctorante*	Université de Tlemcen	Invitée

Année universitaire 2018-2019

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABOU BEKR BELKAID
FACULTÉ DE TECHNOLOGIE
DÉPARTEMENT DE GÉNIE BIOMÉDICAL

MÉMOIRE DE FIN D'ÉTUDES

pour obtenir le grade de

MASTER EN GÉNIE BIOMÉDICAL

Spécialité : **Informatique Biomédicale**

présenté et soutenu publiquement
par

BOUKHOBZA Khadidja

le 26 Juin 2019

Titre:

A Variable Importance Measure for a Cost Sensitive Random forest Predition on a Budget.

Jury

Président du jury.	Pr. CHIKH Mohamed Amine,	UABB Tlemcen
Examinateur.	Dr. EL HABIB DAHO Mostafa,	MCB UABB Tlemcen
Invités d'honneur.	Melle. GUILAL Rima,	Doctorante UABB Tlemcen
Encadreur.	Dr. SETTOUTI Nesma,	MCA UABB Tlemcen
Co-Encadreur.	Dr. SAIDI Meryem,	MCB ESM Tlemcen

To my parents,

Remerciments

Tout d'abord, je voudrais remercier Allah, le Tout-puissant, de m'avoir octroyé la force pour accomplir ce travail et de me bénir en étant entourée par des personnes généreuses qui ont été ma plus grande richesse et mon plus fort support.

Je voudrais saisir cette occasion pour exprimer ma plus profonde gratitude et reconnaissance à mon encadreur, Dr. SETTOUTI Nesma, pour son dévouement, son soutien continu, sa motivation, son enthousiasme et son vigilance tout au long de ce projet.

Mes sincères remerciements vont également à ma co-encadreur, le Dr SAIDI Meryem, pour son soutien, ses précieux conseils et ses encouragements.

Je remercier aussi tous les membres du jury pour leur intérêt pour ce travail et pour avoir pris le temps d'évaluer cette thèse.

Je ne peux pas finir sans dire à quel point je suis reconnaissante à mes parents, mes deux frères: Amine et Mohamed, mes soeurs: Fatima Elzahra, Souad, Kaouther, et la petite Raoudha; qui ont toujours cru en moi et qui m'ont toujours soutenu et encouragé pour faire de mon mieux dans tous les domaines de la vie, sans leur soutien et leur assistance, sans leur amour, leur aide et leur encouragement la réalisation de ce travail n'aurait pas été possible.

J'aimerai aussi remercier toutes les personnes qui ont attendu chaudement le jour de me voir soutenir et qui ont été toujours là pour moi.

Résumé

De nos jours, avoir de bons soins de santé en utilisant moins d'argent devient un défi, car les technologies sont devenues de plus en plus coûteuses et les budgets sont limités. D'autre part, dans le diagnostic médical, une fausse prédiction négative (une personne malade déclarée comme étant saine) peut avoir des conséquences plus graves qu'une fausse prédiction positive et leur attribuer des coûts égaux est inapproprié.
Ce projet de fin d'études contribue à la fois aux apprentissages budgétisés et aux apprentissages sensibles au coût en développant un modèle capable de faire un compromis entre les coûts de classification erronée et les coûts de test. Le modèle proposé est basé sur l'idée d'utiliser les mesures d'importance de variables de la forêt aléatoire en tant que coûts de test et en choisissant l'arbre optimal de la forêt développée en tant que stratégie de test. Notre modèle a été testé sur dix base de données: neuf base de données de UCI Machine Learning et une base de données du monde réel: le myélome multiple; collectée au Centre de Lutte Contre le Cancer (CLCC) de Tlemcen.

Mots clés

Apprentissage sensible au coût, apprentissage budgétisé, forêts aléatoires, mesures d'importance des variables, UCI Machine Learning, Myélome multiple.

Abstract

Nowdays, having a good health-care using less money become a challenge, as technologies became more and more expensive and budgets are limited.On the other hand, in the medical diagnosis, a false negative prediction (a sick person declared as healthy one) may have more serious consequences than a false positive prediction, and assigning them equal costs is probably inappropriate.
This Master thesis makes contribute to both the fields of budgeted-learning, and cost sensitive learning in that it develops a model that can make a compromise between misclassification costs and test costs at the same time. The proposed model is based on the idea of using the variables importance measures of random forest as test costs and choosing the optimal tree from the grown forest as test strategy. Our model has been tested on nine UCI Machine Learning datasets and on a real-world database: multiple myeloma; collected from the anti cancer center of Tlemcen.

Keywords

Cost sensitive learning, budgeted learning, random forests, variables importance measures, UCI Machine Learning Datasets, Multiple myeloma.

الملخص

في الوقت الحاضر اصبح الحصول على رعاية صحية جيدة باستخدام أموال اقل يمثل تحديا، حيث أصبحت التقنيات باهظة الثمن بشكل متزايد و الميزانيات محدودة.
من ناحية أخرى في تشخيص طبي قد يكون للتنبؤ السلبي الخاطئ (شخص مريض يشخص أنه يتمتع بصحة جيدة) عواقب وخيمة اكثر من التنبؤ الإيجابي الخاطئ، ومن المحتمل أن يكون اعتبار تكاليف هذين الاثنين متساوية، غير مضبوط.
تساهم هذه الأطروحة في كل من التعلم المدرج في الميزانية وتعلم الحساس من حيث التكلفة من خلال تطوير نموذج يمكن أن يحدث المفاضلة بين تكاليف تصنيف الخاطئ والتكاليف الاختبار يعتمد النموذج المقترح على فكرة استخدام مقاييس أهمية متغير غابات العشوائية كالتكاليف الاختبار واختيار الشجرة المثلى للغابات المطورة كاستراتيجية اختبار.
تم اختبار نموذجنا على 10 قواعد بيانات :تسع قواعد بيانية من UCI machine learning وقاعدة بيانية في العالم الحقيقي:
الورم النخاعي المتعدد التي تم جمعها في مركز مكافحة سرطان بتلمسان.

الكلمات المفتاحية:

التعلم حساس من حيث التكلفة التعلم في الميزانية الغابة العشوائية مقاييس أهمية متغيرات الورم النخا**عي** المتعدد

Contents

List of Figures

List of Tables

Glossary

ACR: Urine albumin to creatinine ratio.
B2M: Beta-2 microglobulin.
BR: Biased Robin.
BUN: blood urea nitrogen.
CA: chromosomal abnormalities.
CART: Classification And Regression Trees.
CBC: Complete Blood Count.
CostVimp: cost variables importance.
CRP: C-reactive protein.
CS: Cost sensitive.
CSNB: Cost-Sensitive Naïve Bayes algorithm.
CSTree: Cost sensitive tree.
CT: Computerized tomography.
ECG: electrocardiogram.
FDA: Fast Data Acquisition called.
FN: False negative.
FP: False positive.
ICET: Inexpensive Classification with Expensive Tests.
IFE: Immunofixation electrophoresis.
ISS: International Staging System.
LDH: Lactate dehydrogenase.
MDP: Markov Decision Process.
MKL: Multiple Kernel Learning.
MM: multiple myeloma.
MRI: Magnetic resonance imaging.
NN: neural network.
OOB: Out of bag.
PAC: Probably-Approximately-Correct.
PET: Positron emission tomography.
RADIN: Recurrent ADaptive AcquisitionNetwork.
RF: random forest.
RR: Round Robin.
SFL: Single Feature Lookahead.
TP: True positive.
TN: True negative.

Introduction

Machine learning is the science, art and technology of exploring large and complex bodies of data in order to discover useful patterns, classification represents the most important problem in machine learning; classification can be used in a variety of applications, such as medical diagnosis for better automation in healthcare, biological data, object recognition, intrusion detection and many.

Usually, the classic classification problem aims to minimize the number of errors. Neverless, the default assumption of equal misclassification costs is most likely violated; many real-world applications require varying costs for different types of misclassification errors. For instance, a false-negative prediction for a Spam classification system only takes the user an extra second to delete the email, while a false-positive prediction can mean a huge loss when the email actually carries important information; in bacteria classification, misclassifying a Gram-positive species as a Gram-negative one leads to totally ineffective treatments and is hence more serious than misclassifying a Gram-positive species as another Gram-negative one; when classifying a patient as healthy, cold-infected, or Tuberculosis-infected, predicting an Tuberculosis-infected patient as healthy is significantly more serious than predicting a healthy patient as Tuberculosis-infected .

To address this problem, cost-sensitive classification is developed, which considers the varying costs of different misclassification types, a cost-sensitive classification problem can be very different from the regular classification one, and can be used by applications like: targeted marketing, information retrieval, medical decision making, In fact, cost-sensitive classification can be used to express any supervised learning problem.

In medical decision making, there is a sequence of tests that each patient should run to do the diagnostic; theses test are usually expensive and to purchase all these tests we are going to spend some packets which can't be possible for everyone. Therefore, the field of budgeted learning was essentially developed, the biggest challenge of budgeted learning is to find the most informative attributes of each instances to provide the best hypothesis for a model that use the minimal budget.

Thus, the aim of this research is to develop a model that can minimize the misclassification costs and the budget of the diagnostic tests at the same time.
The following section summarizes the organization of this thesis.

- Chapter I: This chapter has presented the Cost sensitive and budgeted learning foundations (including cost types, and structure of learning system).
- Chapter II: This chapter presents the background and a literature review covering the fields of cost-sensitive learning, Budgeted-learning and features selection.
- Chapter III: This chapter introduces the basic notion and algorithms for automatically growing decision trees and random forest, as well as the concept of our proposed algorithm Cost-Vimp.
- Chapter IV: Then in chapter, we analyzed and discussed the performance of our classifier, the evaluation is based on the cost, the budget and the accuracy. It also includes experimentation on synthetic datasets and real-world application: multiple myeloma.

This thesis concludes with a conclusion where we summarize the contribution, review the extent of our objectives, and foresee the future work.

Chapter 1

Cost-Sensitive Learning

Introduction

The most major subset of Artificial intelligence is machine learning it combines techniques which use statistical methods that allow machines to improve with experiences; classification, is an important subject in machine learning and one of the main tasks in knowledge discovery and data mining [3].

In the last years so many effective classification approaches have been developed, such as: naïve Bayes classifier (1968), decision trees (1989), rule induction (1987), discriminant analysis (1975), neural networks (1943), and support vector machines (1995), among many others; their common aim is to generate classifiers that can recognize classes or predict future examples from the labeled discrete or continuous data. Most of those algorithms crave to minimize the error rate as it is the most used measure of the performance of a classifier but they suppose that all errors have equal costs, they do not take into consideration the difference between types of misclassification errors: **cost** in this case is used as a synonym for **loss**.

For example, in the case of the binary classification, false positives (an example is improperly reported as positive "presence of disease") and false negatives (positive example misclassified) may have the same cost. However, this supposition is not true in real-world; some mistakes are just more costly than others.In this context, Cost-Sensitive Learning seems to be a better option.

1 Cost-Sensitive Learning

Cost-Sensitive Learning is a type of learning that takes the misclassification costs (and possibly other types of cost: costs of testing, costs of obtaining data . . .) into consideration, its aim is to minimize the total costs.

Most of classification algorithms totally ignore the cost of misclassification that could be incurred, they suppose that misclassification cost is the same for all instances,that said cost-insensitive learning.

Unlike the cost-insensitive learning, the cost-sensitive learning treats different misclassification differently. That is, the cost of labeling a positive example as negative can be different from the cost of labeling a negative example as positive. Cost-insensitive learning does not take misclassification costs into consideration [4].

2 Types of costs

According to Turney (2000) there are several types of costs that are involved in classification problems. In the literature, misclassification costs are highlighted as being the most important costs in data mining and machine-learning. Costs can be measured in many distinct units as, for instance, money (euros, dollars), time (seconds, minutes) or other types of measures (e.g., quality of life in medical diagnosis).

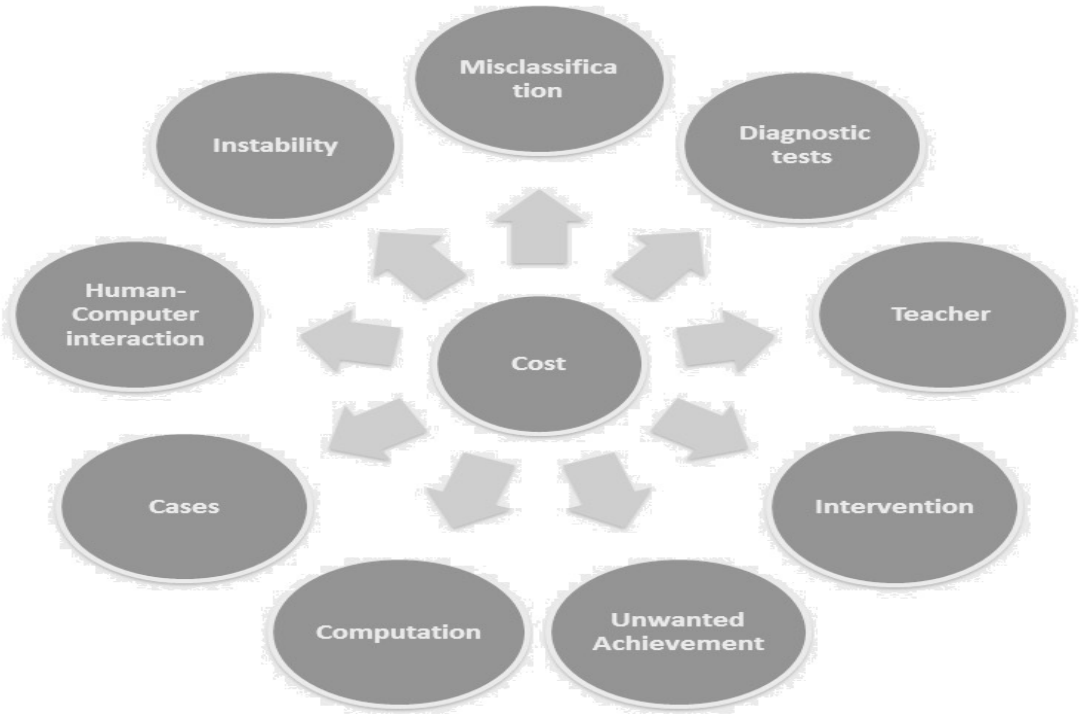

Figure 1.1: Types of cost

- **Cost of misclassification errors:** in some cases, certain types of errors may have the same cost so minimizing the cost is equivalent to minimizing the error rate ,others the cost depend to features , so it is more important to minimize the cost of misclassified examples than to minimize the number of misclassified examples.

- **Cost of diagnostic tests:** Each medical test (obstetric echography or blood test) may have an associated cost. In general, we talk about cost of diagnos-

tic tests only when the cost of misclassification errors is harmless.

- **Cost of classifying cases or Cost of Teacher:** Ask an expert to classify unlabeled examples or to verify the difficult cases has a cost. Active learning is the act of asking the teacher to classify unlabeled examples already selected from a set.

- **Cost of intervention:** it is the cost associated to the effort required to manipulate the process in order to rise or reduce the feature's value.

- **Cost of Unwanted Achievements:** if we mess on the process of intervention we will end with a misclassification error rate increased.

- **Cost of computation:** The size complexity of a computer program , time complexity, space complexity, training or testing complexity are various form of computational complexity and they all have a cost to take into account .

- **Cost of cases:** in Machine learning and data mining acquiring new cases is usually very expensive or almost impossible.

- **Cost of human-computer interaction:** even the best program learner cannot work by it-self, and this interaction with the expert to prepare data, to select predictors, to define parameters or to evaluate a model has a cost.

- **Cost of instability:** it is suitable to have a model that produce relatively close results, the more our model is stable the more we benefit the less is the cost.

From the enumeration above, we can say that the cost of misclassification errors is the most important type; it has a unique position in the taxonomy of Turney and a majority of the machine learning reviews. Unlike the other forms of cost that can be only evaluated in the context of the misclassification error cost. [4] .

In this study, we are going to be more interested about the cost of diagnostic test or what we call too **budgeted learning**.

3 Budgeted learning

When the learning algorithm has free access to the training set class labels but have to pay for using each feature to learn a hypothesis is what we call budgeted

learning , the idea behind budgeted learning is to use the least costly features as much as possible to decrease the average classification cost [5].

The diagnostic of a patient is based on a sequence of tests (features), those medical tests are usually so expensive, our aim is to **correctly** diagnostic a patient while respecting a given budget.

The Costs of Misclassification

Being operated for a non-existed tumor or being untreated for an existed one ? Which mistake would be worse?

The misclassification cost plays its essential role in various cost-sensitive learning algorithms [6].

In cost-sensitive learning, the costs of false positive(actual negative but predicted as positive), false negative (FN), true positive (TP), and true negative (TN) can be given in a cost matrix, the notation C (A, B) is used to represent the misclassification cost of classifying an instance from its actual class A into the predicted class B (1 is used for positive, and 0 for negative).These misclassification cost values can be given by domain experts, or learned via other approaches [7].

The cost of misclassification can be very damaging to patients because allowing an unhealthy person to go untreated can be fatal or have severe side effects.

CONFUSION MATRIX		
	Predicted as positive	Predicted as negative
Actually positive	True positives (TP)	False negatives (FN)
Actually negative	False positive (FP)	True negatives (TN)

Table 1.1: Confusion matrix

COST MATRIX		
	Predicted as positive	Predicted as negative
Actually positive	C (1, 1) (TP)	C (0, 1) (FN)
Actually negative	C (1, 0) (FP)	C (0, 0)(TN)

Table 1.2: Cost matrix
[7]

Costs of Diagnostic Tests

We can talk about the cost of diagnostic test from more than one perspective, running a medical test has a cost either a monetary one (medical test are usually expensive) or considering the time wasted or the quality of life of the patient.

Most of patients turn down a test that could help them because they cannot afford it; they tell themselves that they might not have this disease, so they are not willing to spend extra packets on it because there's not as much risk.

Not only the financial side, but also medical tests can also be risky (Spinal biopsy) or uncomfortable (Fiberoptic).

Overall, diagnostic tests should not be ordered if their costs are greater than the costs of misclassification, if a test is more costly than misclassification errors; it is pointless to run it. On the other hand, if the cost for a set of tests is less than the cost of misclassification errors, it is rational to order all possible relevant tests. These aspects should similarly be considered in a strategy for cost-sensitive learning [4].

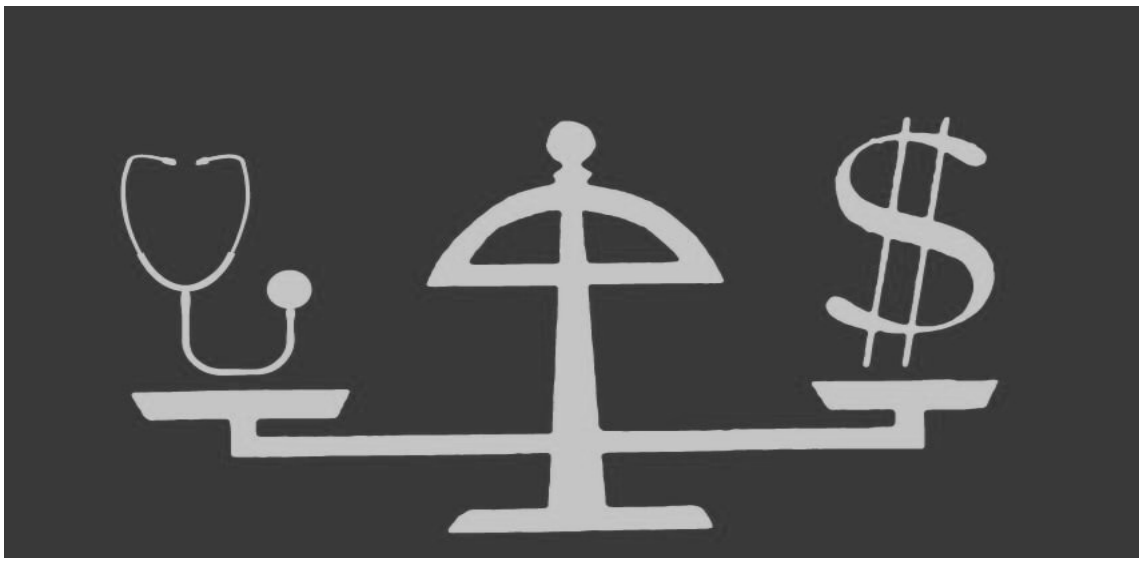

Figure 1.2: Cost of diagnostic test Vs quality of life [1]

4 Structure of Learning System

Depending on the way a cost sensitive algorithm integrates costs it can be categorized into direct methods or cost-sensitive meta-learning methods, the first approach is to build directly a classifier that is cost-sensitive in itself. While the second one, known as the meta-learning or indirect method is to design a wrapper that converts cost-insensitive called also cost-blind learners into cost-sensitive.

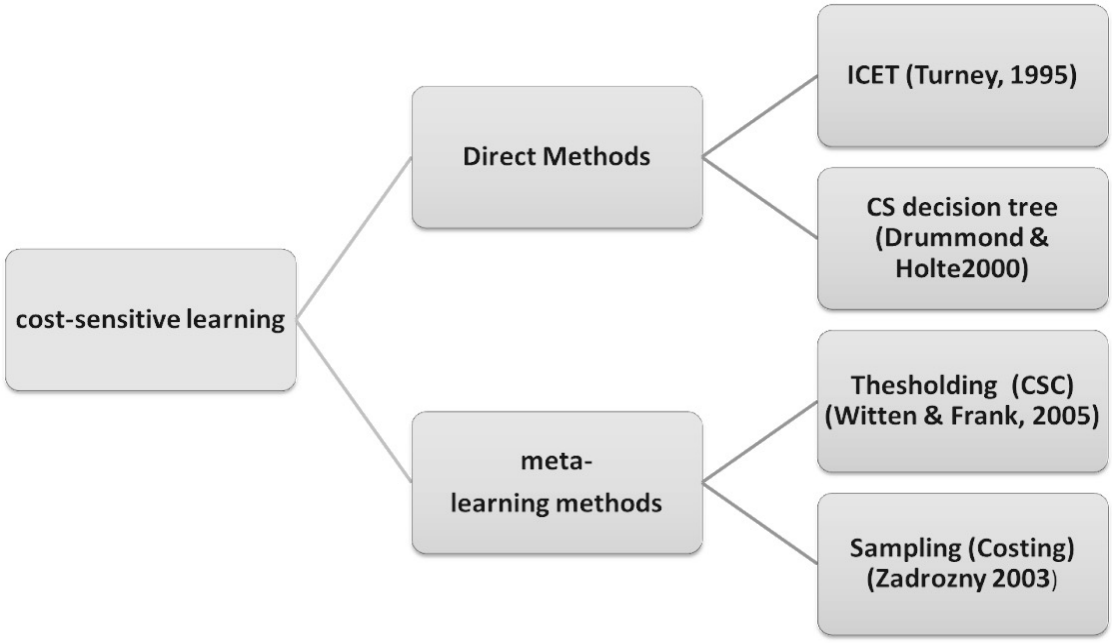

Figure 1.3: Structure of Learning System

- **Direct Cost-Sensitive Learning:** The idea of direct Cost-Sensitive Learning is straightforward; it consists in integrating direct costs of misclassification (or other types of cost) at the learning process. Several works covered this category, such as ICET [8], [7], [9].

  In his work, Turney propose an algorithm that integrates the cost of misclassification in the fitness function of genetic algorithms. While, [7], uses the misclassification costs directly in building process of what called CSTree. Unlike classical decision trees that use Gini, Entropy or accuracy, as criteria to select the best attribute for the construction process, the CSTree take the classification errors into account and selects the attribute as a root of a subtree, that can lead to the minimal total cost.

  In their paper [9], Lomax & Yedara exploit the threshold to adjust the theoretic measure of information, based on the classes's costs in order to include the misclassification costs.

- **Cost-Sensitive Meta-Learning:** Converting an existing cost-insensitive classifier into cost-sensitive one is what called cost-Sensitive Meta-Learning, it is known as a wrapper or a black box that deal with the algorithm as a closed box without modifying any behaviors or parameters of the classifier. The cost-sensitive Meta learning itself can be categorized into two grades: Algorithms that use **thresholding** and those that use **sampling**.

  1. Thresholding: We can transform a cost-insensitive classifier into cost sensitive one by simply choosing a threshold to classify examples into

positive or negative if this one can produce accurate probability estimation.

2. *MetaCost: (Relabeling)* as a first step this algorithm uses the bagging approach. The main idea is to relabel training instances with an optimal class according to the minimal cost, and then building new classifier that can predict the label of test instances. This is known as sampling with labeling [10].

3. Sampling: This algorithm changes the occurrence of instances in the training set according to the cost of label of each one (increases the number of the costly instances: Over-sampling; reduce the number of the less costly ones: Under-sampling). Sampling is an optimal solution for the problem of imbalanced data because it does not change the algorithm itself but just arranges the distribution of data in order to be more rational toward the costly classes [6]. Two approaches are proposed to apply sampling: random sampling and determinate sampling. [11]

4. *Weighting:* In his paper [12] ,Ting proposes to associate each instance with a weight according to its cost, in such way weights and costs are proportionally tied (the greater is the cost of misclassifying an instance the higher is its weights).

5. *Costing:* (rejection sampling) Sampling techniques (duplicating instances in the training set), may produce over-fitting in the model construction, to avoid that [13] proposes costing : keep all instances of the rare class and sampling those of the majority class without replacement according to the cost of each instance, then applying bagging in order to minimize the misclassification cost.

5 Conclusion

We presented in this chapter a brief overview of Cost sensitive learning and its methods; In our study, we are going to be more interested at using a Meta-sensitive learning method as this approach opts to enrich the data instead of modifying the classifier's parameters.

Chapter 2

Current cost-sensitive and budgeted learning approaches

Introduction

Most of the classification algorithms in the literature used the error rate to evaluate the performance of a classifier, so minimizing this rate means eventually a good classifier. Some other works take into account non-uniform misclassification costs, that is, different costs for different types of errors; however, in real world problems there are different costs for different types of errors. The cost sensitive learning takes into account the non-uniform misclassification costs. The cost sensitive learning is also used to address the class imbalanced problem (when, in a classification problem, there are many more instances of some classes than others) [14]. Increasing the cost of misclassifying minority classes can minimize the imbalanced class problem. Re-sampling the training set is also a way to build an algorithm sensitive to costs.

Other research claimed that it is interesting to take the cost of tests into consideration, but those works ignore misclassification costs while some other works are concerned simultaneously with several types of costs. On the other hand, some works tried to tackle the Cost-Sensitive Feature Acquisition problem.

1 Overview of Cost sensitive approaches

In the previous chapter, we talked about the several types of cost and we said that the misclassification and test cost are the most important two; we precise that we cannot talk about the other types of cost unless the cost of misclassification is harmless. In the next section, we are going to present the most known papers in literature that consider misclassification and test cost, we can distinguish four categories for the cost sensitive problem : manipulating data to get Cost-Sensitive learner, Cost Sensitive Trees, Naïve Bayes Classifiers and those who consider cost sensitive problem as a reinforcement Learning problem and use the Markov Decision Processes.

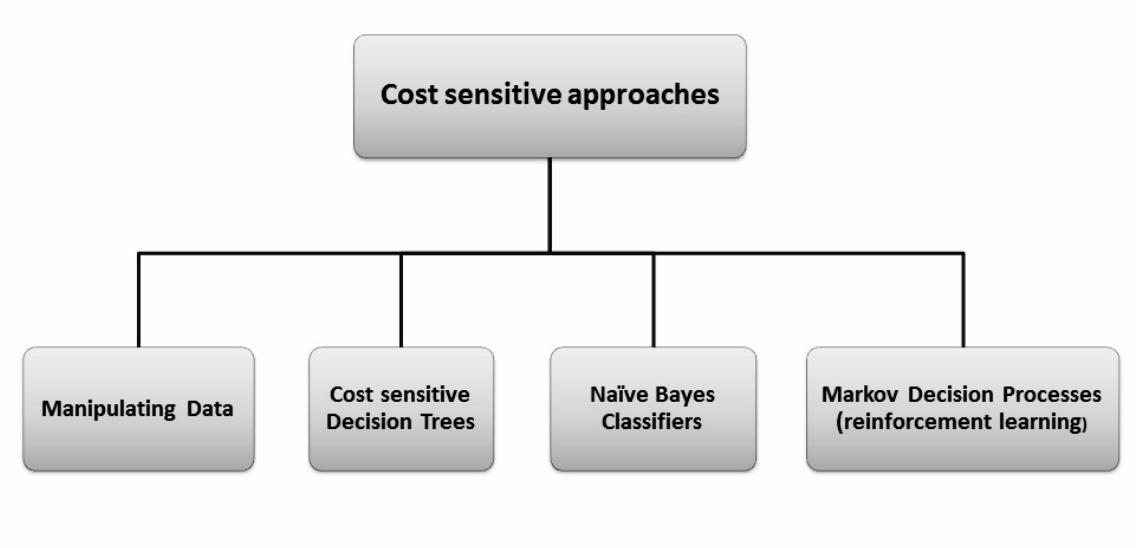

Figure 2.1: Cost sensitive approaches

1.1 Manipulating Data to Obtain Cost-Sensitive learner

A meta-classifier is known as an Algorithm that can manipulate training set or its outputs in order to obtain cost-sensitive classifiers. One approach is to change the class distribution in order to minimize, the costs of new instances. These changes aim to give each class a distribution proportional to its importance (increasing minority class and minimizing majority one). This process is known by *under-sampling* or *oversampling* . [14].
Another approach is *MetaCost* as a first step the algorithm uses the bagging approach. The main idea is to relabel training instances with an optimal class according to the minimal cost, and then building new classifier that can predict the label of test instances. This is also known as sampling with labeling [10].
Another approach, without using sampling,*threshold adjusting* choosing a threshold to classify examples into positive or negative ones [6].

1.2 Cost sensitive Decision Trees

1. *Decision Tree Optimized by a Genetic Algorithm:* The work of Turney [8] is known to be the first that consider both test and misclassification costs, he implemented a system called ICET (Inexpensive Classification with Expensive Tests), which build a decision tree using a genetic algorithm that minimizes test and misclassification costs at the same time. The ICET system was robust but very time consuming. Turney considered that his method for the cost-sensitive classification problem was, basically, a reinforcement learning problem.

   Later, [15] showed that in a preprocessing phase the efficiency of learning can be significantly improved by removing irrelevant attributes. The cost-sensitive elimination of attributes improved the learning efficiency of the

hybrid algorithm ICET.

2. *Cost-Sensitive Decision Trees:* In their paper Ling et al., [16] come with another approach to build and test cost sensitive decision trees (CSTree). This approach was sensitive for both types of costs (the misclassification and test cost), the proposed algorithm used a new splitting criterion to select the node parent attribute, and it chooses the attribute that minimizes the total cost, instead of minimal entropy (as in C4.5).

3. *Specific Decision Trees and Hybrid Approaches:* Another approach for learning cost-sensitive decision trees was proposed by [17]. Instead of building a single decision tree for all test examples, the proposed method builds a different tree for each new test example with a different set of unknown attributes. This process considers costs only for attributes with unknown value (the test cost of known attributes is 0).

   In another paper, Sheng & Ling [18] proposed a hybrid cost-sensitive decision tree to reduce the minimum total cost. The proposed model integrates cost-sensitive decision trees (to collect required tests) with cost-sensitive naïve Bayes.

   Later, Sheng et al. [19] updated their strategy to build decision trees sensitive to costs, with the insertion of three medical test strategies, sequential test, single batch and multiple batch tests, to decide and order witch attributes to run tests on.

4. *Taking Risk into Account:* Freitas et al. [20] presented an approach to combine several types of costs with relevance for health management. They defined algorithm for the induction of cost-sensitive decision trees, including misclassification costs, costs associated with risk, delayed costs and test costs. This approach used different strategies to test models, including group costs, common costs, and individual costs. The aim was to build decision trees that minimize the costs and be the most “patient-friendly”, penalties was integrated for risky tests invasive or delayed tests.

1.3 Markov Decision Processes

Some authors as in [21] considered the problem of cost-sensitive learning as a Markov Decision Process (MDP) that has the disadvantage of being computationally expensive. They adopt an optimal search strategy (heuristic AO* algorithm "Nilsson 1980"), which may incur a high computational cost and be very time consuming. To overcome this problem some authors [22], [23] propose to integrate the notion of time (a cost for the time passed).
Arnt & Zilberstein [22], involved a cost for the time needed to obtain the result of a test, in which they considered attribute (test) costs, misclassification costs and

also a utility cost related to the time passed while measuring attributes. As in the work of [21], they also modeled the problem as a Markov Decision Process and then used the search heuristic AO*. They tried to compromise between time and accuracy, and proposed an approach to attribute measurement and classification for a variety of time sensitive applications.

Sheng et al. [23] had seen that in most of real world application data are not available and getting it is usually time and money costly so he proposed an on-line framework for Fast Data Acquisition called FDA, this system can estimate the number of examples needed in each acquisition and acquire them simultaneously. Comparing to the naïve step-by-step data acquisition strategy, FDA reduces significantly the number of times of data acquisition and model building.

1.4 Naïve Bayes Classifiers

Greiner et al. [24] studied the problem of active learning classifiers basing on a variation of the Probably-Approximately-Correct (PAC) model, they proposed a learning and active classification framework that show how to use a budget in collecting the pertinent information for applications with no actual data at beginning. The learner "pays" to see any attributes (learning costs) and has to predict the classification for each instance, with possible penalties.

Lizotte et al. [5] studied an active learning situation where the classifier (naïve Bayes), with a hard budget, could "buy" data during training. Considering that each attribute of a training data has an associated cost, and the total cost during training must remain less than the fixed budget. They compared methods for sequentially deciding which attribute value to purchase next, considering budget limitations and knowledge about some parameters of the naïve Bayes model.

Chai et al. [7] proposed a Cost-Sensitive Naïve Bayes algorithm, called CSNB that can reduce the total cost of attributes and misclassification at the same time in this paper they integrate sequential and batch test strategies to determine which feature is selected to be "purchased" (tested).

Leveling et al. [25] proposed a formal justification for a decision function under the Bayesian decision framework that comprises the minimization of Bayesian risk and an empirical decision function.

This section surveyed the cost sensitive problem, we have seen approaches that tackle this problem and methods that take cost into consideration, and the Table 2.1 below summarizes some of recent work in the field of cost sensitive learning.

AUTHORS	TITLE	ALGORITHM	APPLICATION	TYPE OF COST
Sun et al. 2007 [26]	Cost-sensitive boosting for classification of imbalanced data	AdaBoost : AdaC1-AdaC2-AdaC3-AdaCost-CSB2	Breast cancer, Hepatitis, Pima Indian's diabetes database (Pima), and Sick-thyroid from UCI datasets	Misclassification costs
Weiss et al. 2007 [27]	Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?	Oversampling and Under-sampling techniques	Synthetic and real-world benchmark	Uniform & non-uniform misclassification costs
Lev Reyzin. 2011 [28]	Boosting on a Budget: Sampling for Feature-Efficient Prediction	AdaBoost for Uniform and non-uniform costs	UCI datasets: census, splice, ocr17, and ocr49	Uniform and arbitrary feature costs
He et al. 2012 [29]	Cost-sensitive Dynamic Feature Selection	DAgger for Feature Selection	Radar signal (binary), digit recognition (10 classes) and image segmentation (7 classes)	feature cost on test-time.
Xu et al. 2012 [30]	The Greedy Miser: Learning under Test-time Budgets	Greedy Miser	The Yahoo Learning to Rank Challenge data set the scene recognition data set	Feature extraction cost.

Karayev and al.2013 [31]	Dynamic Feature Selection for Classification on a Budget	Gaussian Naive Bayes	Imagenet subset and Scenes-15 dataset	Features costs with fixed budget
Ma et al. 2017 [32]	CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests	random forests CURE-SMOTE	UCI datasets	Oversampling and misclassification cost

Table 2.1: Summary of recent Cost sensitive works

2 Overview of budgeted learning approaches

If the goal of machine learning in general is learn to predict than we can say that budgeted learning answer the question *what to learn?*

The biggest challenge of budgeted learning is to find the most informative attributes of each instances to provide the best hypothesis for a model that use the minimal budget. If we want to simply distinguish between active and budgeted learning we can say that in the first one we don't have a fixed budget for each example while in budgeted learning we have a hard budget which the model should respect.

2.1 Budgeted feature selection and acquisition approaches

The aim is either to reduce the number of the necessary used features when they all have the same cost "price" or to minimize the total cost when they have non uniform costs (medical test).

For that several approaches were suggested; starting by *feature selection methods* that propose to limit the set of features being used for training ([24]; [33]); but those methods lack the aspect of adaptability and the same set is used for all inputs. Then, [34] proposed *probabilistic methods* that can measure the information value of each feature based on the current evidence; however those methods are greedy and computationally expensive when applied on large dataset. Some authors later, ([35]; [30]) suggested as an intuitive way to integrate *feature costs*.

Another strategy was proposed by [36], where they use greedy min-max on random forest algorithm to integrate feature costs. More recently, [37] proposed an approach that employs adaptive linear or tree based classifiers, alternating between low-cost models for easy-cost to handle instances and higher-cost models to handle more complicated cases.

More recently, a new Axe appear *Adaptive cost-sensitive feature acquisition*, the aim is to develop a new family of models able to acquire information by themselves (information needs to be acquired), to choose what to compute (different computations are applied to different inputs) and to handle operational constraints (size of data), it is about *what* to learn *when* to learn, *how much* to learn. In this field G. Contardo [38] propose a system his aim is to learn to actively learn, he define a model based on NN called RADIN (Recurrent ADaptive Acquisition Network) that considers all examples of a dataset before predicting which example should be labeled. Kachuee et al. [39] proposed a method based on deep Q-networks for cost-sensitive feature acquisition at the prediction time. The proposed solution employs uncertainty analysis in neural network classifiers as a measure for finding the value of each feature given a context.

2.2 Selection of instances considering budget

Other authors saw that selecting which features to purchase is not enough (selecting attributes to test and then choose randomly an instance) they opt to select instances and attributes, choosing features that minimize the total cost and instances that are more susceptible to be misclassified.

Uniform sampling and Error Sampling those two methods were proposed by [40] and [41] to consider only a part of instances instead of all of them. In this method, they apply the sampling then choose an (instance, feature) pair.

M. Saar-Tsechansky [42] proposed to use Log Gain instead of conditional entropy or GINI index to measure the importance of the feature to select. In parallel they also tried to reduce the search space of instances by choosing only those ones that are wrongly predicted. Deng et al. [43] present new heuristics that can select an instance to purchase after the attribute is selected, instead of selecting an instance randomly.

In this section, we presented learning on a budget or budgeted learning approaches and the Table 2.2 below presents some of recent works in this field.

AUTHORS	TITLE	ALGORITHM	APPLICATION	TYPE OF COST
Lizotte et al. 2003 [5]	Budgeted Learning, Part I: The Multi-Armed Bandit Case	Round Robin (RR) and Random-Greedy Algorithms	Budgeted multi-armed bandit problem.	Feature costs

Kapoor et al. 2005 [44]	Budgeted Learning of Bounded Active Classifiers	Optimal Policy, Round Robin (RR), Biased Robin (BR), Single Feature Lookahead (SFL) and Randomized SFL	Synthetic and real-world benchmark	Feature costs with fixed budget
Guha et al. 2007 [45]	Approximation Algorithms for Budgeted Learning Problems	Approximation algorithms and Greedy Order	Budgeted multi-armed bandit problem.	Feature and instances costs
Bontempi et al. 2011 [46]	A Selecting-the-Best Method for Budgeted Model Selection	a variation of Monte Carlo stochastic approximation	Synthetic and real-world benchmark	Feature and instances costs
Yang et al. 2015 [47]	Budget Constrained Non-Monotonic Feature Selection	Multiple Kernel Learning (MKL)	Synthetic and real-world benchmark	Cost on the feature subset size
Nan et al. 2015 [48]	Feature-Budgeted Random Forest	Budget random forest and Greedy Miser	4 real world benchmarked datasets	Feature costs
Nushi et al. 2016 [49]	Learning and Feature Selection under Budget Constraints in Crowd sourcing	B-LEAFS	synthetic and real-world crowd sourcing data	Feature costs on training and test phase
Nan et al. 2016 [50]	Pruning Random Forests for Prediction on a Budget	random forest (RF)	four benchmark datasets	Feature cost and error-cost trade-off .

Shim et al. 2018 [51]	Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding	Markov decision process (MDP)	Synthetic dataset	feature acquisition cost

Table 2.2: Budgeted feature selection and acquisition approaches

3 Conclusion

To conclude it is important to say that the cost sensitive problem attracted so many researchers and for years they tried to tackle it from more than one perspective. The real motivation behind all those works come from the fact that there is no best solution or optimal algorithm for all the problems and in this context come our study trying to find a compromise between misclassification and test costs for medical data.

Chapter 3

Proposition and Methods

Introduction

In many real-world tasks, it is well known that an ensemble is usually significantly more accurate and can achieve great success, so it is straightforward that Ensemble methods techniques in machine learning outperform single classifiers.

This kind of state-of-the art learning approach has been widely studied in the few last years. The main idea of ensemble methods is to randomize the learning procedure in order to generate different classifiers from a single learning set, and then combine those basic classifiers to perform the final prediction. In order to induce the random permutations, several methods have been proposed, in particular: bagging (1996) [52], pasting (1999) [53], random forests (2001) [54] and random patches (2012). Finally, after the base classiers are trained, they are typically combined using either majority voting; Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes [55], weighted voting: Unlike majority voting, where each model has the same rights, we can increase the importance of one or more models [55] or stacking: it is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features [55].

According to Dietterich [56] there are three main reasons why ensemble methods perform better than single models:

1. *Statistical issue:* It is often the case when the learning set is too small and the hypothesis space is too large, the learning algorithm may find several different models with the same performance on the training data. Combining all these models, can reduce the risk of choosing the wrong model.

2. *Computational issue:* In general, learning algorithms rely on some local search optimization and may get stuck in local optima. Then, an ensemble may solve this by running the local search from many different startings across the training data.

3. *Representational issue:* In many machine learning tasks, the true function f cannot be represented by any of the candidate hypotheses. By combing several hypotheses in an ensemble, it may be possible to obtain a model that can expand the space of representable functions.

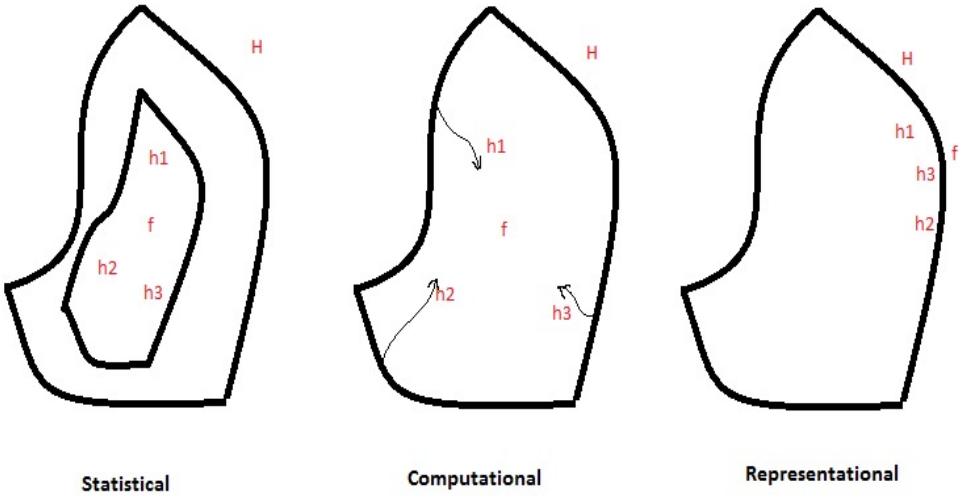

Figure 3.1: Reasons why ensemble methods perform better than single models
A learning algorithm can be viewed as searching a space H of hypotheses to identify the best hypothesis in the space.The point f is the true hypothesis, and we can see (right) that by averaging the accurate hypotheses, we can find a good approximation to f; this is the statistical reason
An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual classiers, as shown in the (bottom)of the figure
By forming weighted sums of hypotheses drawn from H , it may be possible to expand the space of representable functions. Figure 3.1 depicts this situation.

For classification or regression problems, Random Forests (RF) [54] are very hard to beat in terms of performance. Of course we can probably always find a model that can perform better, like a neural network, but these usually take much more time in the development, unlike Random Forests (mostly fast).

On top of that, it provides a pretty good indicator of the importance assigned to features. Although it has its limitations,the RF algorithm is a simple and flexible tool. It's hard to build a "bad" Random Forest, because of its simplicity.

For all that the RF algorithm seems to us to be a great choice to tackle the cost sensitive problem in this work.

1 Random Forest Algorithm

A Random Forest (RF) is a combination of Bagging [52] and Random Subspace [57], consisting of many binary or multi-way decision trees. The final decision is

made by majority voting to aggregate the predictions of all the decision trees. As show 3.2 the random forst procedure :

1. First, training sets are constructed by using a bootstrap mechanism randomly with replacement.
2. Random features are selected with non-replacement from the total features when the nodes of the trees are split.
3. For each subensemble a decision tree is constructed, by calculating the daughter nodes using the same best split approach until the trees are formed with a root node and having the outcome as the leaf node.
4. Finally, outcomes are gathered from all trees, then considering the high voted predicted outcome as the final prediction for the random forest algorithm.This concept is known as majority voting.

The size of the feature subset is usually far less than the size of the total features. The resulting classifiers are robust, very easy to train, accurate, and yield strong performance [54].

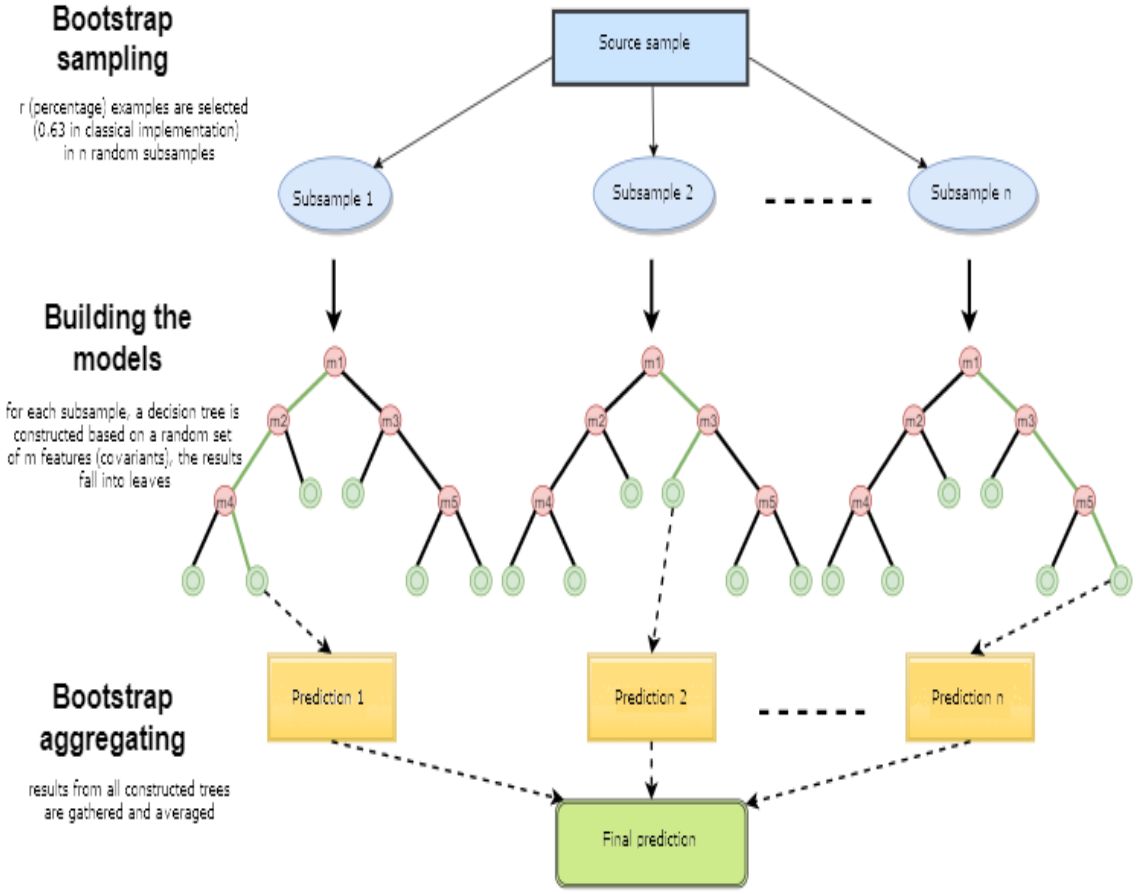

Figure 3.2: The RF classification procedure [2]

L. Breiman [54] propose Bagging and Randomization technique to grow many classification trees with the largest extent possible without pruning. Random Forest is especially attractive for following reasons:

- First, real-world data is usually noisy and can contain many missing values, RF has an effective method for estimating missing data and can maintain good accuracy when a large proportion of the data are missing.
- Furthermore, it has methods for balancing error in class population unbalanced data sets.
- RF can handle a lot of different feature types, like binary, categorical and numerical.
- Random forest can generate an internal estimate of the generalization error as the forest building progresses.
- The RF algorithm gives estimates of what variables are important in the classification and give information about the relation between the variables and the classification.
- RF can also offer an experimental method for detecting variable interactions.
- RF show high predictive accuracy and are applicable in high-dimensional problems with highly correlated features, especially in the situation which often occurs in bio-informatics, like medical diagnosis.
- The capabilities of RF can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.

1.1 Decision trees

One of the most successful ensemble learners is random forests (RF), as their name suggest, the random forest algorithm creates the forest with a number of trees. In general, in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.To understand the random forest model, we must first learn about the decision tree, the basic building block of a random forest.
Decision trees are one of the most promising and popular machine learning algorithms [58]. This technique is considered as a white box: so easy to interpret, has a very low computational cost, and can maintain a good performance compared with other complex techniques [59].

Depending on the impurity measure used during the split, we can distinguish two main categories of decision trees. First the CART algorithm that is based in the Gini index, and later the ID 3 and C 4. 5 which uses the entropy measure.
Both Gini and entropy are measures of impurity of a node. A node having multiple classes is impure whereas a node having only one class is pure. Entropy in statistics is analogous to entropy in thermodynamics where it signifies disorder.

If there are multiple classes in a node, there is disorder in that node.
Information gain is the entropy of parent node minus sum of weighted entropies of child nodes.
Weight of a child node is number of samples in the node/total samples of all child nodes. Similarly information gain is calculated with Gini score.

$$Gini = 1 - \sum_{i=1}^{n} p^2(C_i)$$

$$Entopy = \sum_{i=1} -p(C_i)\log_2(C_i)$$

Where $p(C_i)$ is the probability/percentage of class C_i is a node.

1. CART or Classification And Regression Trees were introduced by Brieman [60]. It is based on using the Gini index as the impurity measure and the tree is grow until all examples in each leaf belong to the same class. Afterwards, the tree is pruned using the cost-complexity method [61].

2. ID3 algorithm uses entropy as the impurity measure. The growing of the tree stop when all examples belong of each leaf belongs to the same class. In ID 3 no pruning is applied [62].

3. C4.5 the extension of ID3 both proposed by Quinlan [62]. Both are similar regarding the measure used, but C 4. 5 define the stopping criteria during the growth process to be when the number of examples in a set is less than a threshold. Moreover, after the tree is created an error based pruning is applied [61].

In his important paper L. Breiman [54], grow an ensemble of CART trees using the Bagging techniques and let them vote for the most popular class, he calls these procedures random forests.

1.2 Classification rules and algorithmic procedure

The best attribute can be computed by three methods: information gain, information gain rate and Gini coefficient, which correspond to ID3, C4.5 and CART, respectively. When the attribute value is continuous, the best split point must be selected.

There are several ways by which the termination criteria for RF can be met:

- Termination occurs when the decision tree reaches maximum depth,
- The impurity of the end node reaches the threshold,
- The number of final samples reaches a set point,
- The candidate attribute is used up.

The RF classification algorithm and procedure are shown below1.

Algorithm 1 The RF Algorithm

Input: training set, testing set, $nTree$: tree number, k: hyper parameter, N: size of subensemble, attribute select method, termination criteria
Output: RF classification model and classification results.
For i=1:$nTree$
-Use the bootstrap method to produce training sets with size N for each tree,
-Select k attributes randomly building nodes and split the dataset by the best attribute,
-Generate each tree recursively without pruning
End
Calculate the probability of unknown sample x belonging to class c,

$$P(c|x) = (1/nTree)\sum(h_j(c|x));$$

Return Predict class through majority voting and calculate OOB error;

$$C \leftarrow \arg maxP(c|x);$$

1.3 Random Forest Variable Importance Measure

Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction.

There are two measures of importance given for each variable in the random forest. The first measure is based on how much the accuracy decreases when the variable is excluded. This is further broken down by outcome class. The second measure is based on the decrease of Gini impurity when a variable is chosen to split a node.

- **Accuracy-based importance:** Each tree has its own out-of-bag sample of data that was not used during construction. This sample is used to calculate importance of a specific variable. First, the prediction accuracy on the out-of-bag sample is measured. Then, the values of the variable in the out-of-bag-sample are randomly shuffled, keeping all other variables the same. Finally, the decrease in prediction accuracy on the shuffled data is measured.The mean decrease in accuracy across all trees is reported. This importance measure is also broken down by outcome class.

- **Gini-based importance:** When a tree is built, the decision about which variable to split at each node uses a calculation of the Gini impurity.For each variable, the sum of the Gini decrease across every tree of the forest is accumulated every time that variable is chosen to split a node. The sum is divided by the number of trees in the forest to give an average. The scale is irrelevant: only the relative values matter.

Neither measure is perfect, but viewing both together allows a comparison of the importance ranking of all variables across both measures.Through looking at the feature importance, we can decide which features we may want to drop, because they don't contribute enough or nothing to the prediction processes. This is important, because a general rule in machine learning is that the more features you have, the more likely your model will suffer from overfitting and vice versa.

2 Proposition

In order to address the cost sensitive problem in medical data, we proposed in this work, a Cost Sensitive Random Forest Variable Importance algorithm named *CostVimp Algorithm*, that takes into consideration the misclassification costs. Our proposition works on two steps:

- First, the matrix cost fixed for all of the datasets (Table.3.1) is introduced into the Random forest induction phase.
- Second, we minimize the total budget, by creating new-costs (features costs or diagnostic test cost) based on the importance variables measures generated by the RF Algorithm.

We are interested in measuring how good is our classifier in terms of cost and budget not only in terms of accuracy because minimizing misclassification rate does not lead to the same results than minimizing cost.

It is important to precise that the definition of cost matrix is quite subjective. For example in Table.3.1, labeling a positive instance as negative (FN) is five time more costly than labeling a negative one as positive (FP).

COST MATRIX		
	Predicted as positive	Predicted as negative
Actually positive	0 (TP)	5 (FN)
Actually negative	1 (FP)	0(TN)

Table 3.1: The proposed cost matrix for experiments

Our proposed algorithm can give a compromise between cost and budget, it aim to minimize the misclassification cost and the total budget at the same time. On the top of that our proposed test strategy (choosing the optimal tree), is extremely fast, easy to interpret and make for straightforward visualizations.

In algorithm 2 the pseudo-code of the proposed RF growing procedure is presented.

Algorithm 2 The CostVimp Algorithm

Input:training set, testing set, $nTree$: tree number, N: size of sub-ensemble, attribute select method, termination criteria, matrix cost,
Output: Tree classification model
For i=1:$nTree$
- Use the bootstrap method to generate a sub-ensemble N for each tree;
-Select variables set randomly and split each sub-ensemble by the best attribute;
-Grow each tree recursively without pruning;
-Calculate the total cost of each tree;

$$Total\ cost = \frac{(nbrFP * costFP + nbrFN * costFN)}{(NbrP * costFP + NbrN * costFN)}$$

End
- Calculate the importance for each feature.

$$fi_i = \frac{\sum nodes\ splits\ on\ feature\ i}{\sum allnodes}$$

- Calculate The normalized feature importance for i in tree j.

$$normfi_{ij} = \frac{fi_i}{\sum all\ features fi_{ij}}$$

- Get the importance measures of each variable generated by the training RF,

$$RFfi_i = \frac{\sum_i \epsilon all\ trees\ normfi_{ij}}{nTree}$$

Calculate the total budget of each tree,

$$Total\ budget = \sum (unique\ used\ predictors\ importance\ measures)$$

$$Somme = \sum normfi_{ij}$$

Choose the best tree from the forest, then predict test class;

$$Best\ tree = \min(Total\ cost) \& (Total\ budget < Somme)$$

Return The best Tree classification model

3 Conclusion

To conclude we can say that Random Forest is one of the most machine learning used algorithms, because of its simplicity and the fact that it can be used for both classification and regression problems.
RF is a flexible, easy to use and produce great result most of the time next chapter we are going to test its performance on the cost sensitive task.

Chapter 4

Experiments and Results

Introduction

The most important process in developing a classifier, it involves evaluating the result of applying different datasets. In this chapter, we are going to evaluate the performance of our classifier on ten datasets, nine datasets from UCI Machine learning and one real world database.

In this research, cost, budget and accuracy are used to evaluate the performance of the model. In this section we present the experimental results. Our experiments are going to be in two phases:

Experiment 1 - standard Datasets we perform experiments on UCI's standard medical datasets using the importance variables measures generated by RF as features costs and a fixed budget.

Experiment 2 - Real World Dataset we test our Cost sensitive Random Forest Variable Importance (*CostVimp* Algorithm) on a real world dataset: multiple myeloma, presented in next section with the real test prices.

1 Experiment 1 - standard Datasets

To evaluate the performance of our proposed algorithm, we first evaluate the different trees generated by RF without pruning, by the cost of misclassification ratio (eq.4.1).

$$Total\ cost = \frac{(nbrFP * costFP + nbrFN * costFN)}{(NbrP * costFP + NbrN * costFN)} \tag{4.1}$$

- nbr FP: Number of misclassified positive instances.
- nbr FN: Number of misclassified negative instances.
- NbrP: Number of all positive instances.
- NbrN: Number of all negative instances.

- cost FP: Cost of misclassified positive instances.
- cost FN: Cost of misclassified negative instances.

Next, we try to find a compromise between misclassification cost and total budget (eq.4.2), we are going to choose the optimal tree with the minimal cost of misclassification ratio and which at the same time minimize the total budget.

$$Total\ budget = \sum(unique\ predictors\ costs) \tag{4.2}$$

For the experiments we used nine standard datasets from UCI repository [1] (all the used datasets are binary classes).

Datasets	Variables	Instances	Class distribution
thoracic surgery	16	470	0,14 ; 0,85
EEG Eye	12	14980	0,44 ; 0,55
Pima	8	768	0,65 ; 0,34
Bupa	6	345	0,42 ; 0,57
Cardio	22	129	0,24 ; 0,75
Mammography	5	830	0,48 ;0,51
Breast cancer	9	699	0,65 ; 0,34
Fertility	9	100	0,88 ; 0,12
South Africa HeartD	9	462	0,65 ; 0,34

Table 4.1: Description of the choosen UCI Datasets

1.1 Results and Discussion

For each dataset, we first calculate the class distribution to be able to split successfully the original dataset into training (0.7) and testing (0.3) dataset.

We next use the bootstrap technique on the training set to grow 100 trees with random subsamples considering the cost matrix to minimize the error and the importance variable measure generated by the RF algorithm as test costs.As test strategy, we are going to choose the best tree from the grown forest. Some variables are either irrelevant or have no impact on the learning process so the optimal tree use only the most important variables, it is well known in the machine-learning community that irrelevant variables can have a negative effect on a

[1] https://archive.ics.uci.edu/ml/index.php

learner's predictive power and has some disadvantages, using only the most important variables decrease systematically the total budget and increase the accuracy. The results are shown in Table 4.2, the column of total budget show: the budget of the used predictors in the optimal tree/ the budget of All the predictors.

Datasets	Rnadom Forest		Cost Vimp			
	Error rate	**Miss-classification Cost**	**Error rate**	**Miss-classification Cost**	**Budget**	**Selected variables**
Thoracic surgery	0.20	0.17	0.25	0.12	94.47%	11
EEG Eye	0.25	0.27	0.22	0.44	100%	12
Pima	0.29	0.34	0.31	0.47	100%	8
Bupa	0.44	0.35	0.31	0.36	100%	6
Cardio	0.16	0.10	0.13	0.19	59.33%	4
Mammography	0.24	0.25	0.24	0.45	100%	5
Breast cancer	0.04	0.03	0.04	0.47	99.64%	8
Fertility	0.14	0.14	0.12	0.01	54.01%	3
South Africa HeartD	0.32	0.38	0.29	0.47	94.47%	8

Table 4.2: Results of the proposed CostVimp on nine datasets

It is well known that there is no single algorithm that performs best for all datasets and this is the case of our algorithm too. From the Table 4.2 above we can summarize those notes:

- The error rate decrease as the total budget increases (breast cancer, Fertility).
- The features selection is not really meaningful when the number of features is limited (EEG Eye, Pima, Bupa, Mammography: the grown trees use all the features, so the budget can't be minimized).
- We used the importance variables measures as costs to calculate the budget; however, this assumption is not always true in the real world. We can find an expensive test (MRI Scan) that is complementary for some cases and vice versa.
- The definition of matrix cost was quite subjective, the results would be more accurate if it was given by domain experts, or learned via automatic approaches.

2 Experiment 2 - Real World Dataset on Multiple Myeloma Disease

2.1 Overview of Multiple myeloma

Multiple myeloma is a cancer that forms in a type of white blood cell called a plasma cell. Plasma cells help the body fight infections by making antibodies that recognize and attack germs. Multiple myeloma causes cancer cells to accumulate in the bone marrow, where they crowd out healthy blood cells. Rather than produce helpful antibodies, the cancer cells produce abnormal proteins that can cause complications[2].

The International Staging System (ISS) is the most commonly used for staging the multiple myeloma, the system is based on two important factors, Beta2 microglobulin and Albumin see Table 4.3.

STAGE	CRITERIA
I	Serum Beta2 microglobulin < 3.5 mg/l Serum albumin >= 35 g/dl
II	Not ISS stage I or III
III	Serum Beta2 microglobulin >=5.5 mg/L

Table 4.3: International Staging System (ISS) for multiple myeloma

2.2 Description of the dataset

The MM dataset was collected by R. GUILAL at the Anti-Cancer Center of University hospital of TLEMCEN, Algeria [3]. It consists of 200 patients who are diagnosed during the period 2008-2019, and 57 features including cover demographic information, personnel and family antecedents, different results of medical exams and tests diagnosis of MM.

To be able to perform our experiment on the MM dataset we have selected only 43 features (only diagnostic tests that have a monetary cost) and 149 instances from the three stages of MM (all the instances are pathologicals), the suspicious cases was deleted. The medical signification of each parameter in the dataset of Multiple myeloma are described as follow[4]:

- **Complete Blood Count (CBC):** complete blood count (CBC), also known as full blood count (FBC) or full blood exam (FBE) or blood panel, is a test

[2]https://www.myeloma.org/
[3]http://www.chu-tlemcen.dz/
[4]https://www.myeloma.org/

panel that gives information about the cells in a patient's blood.

- **Bone marrow examination:** it refers to the pathologic analysis of samples of bone marrow obtained by bone marrow biopsy and bone marrow aspiration. Bone marrow examination is used in the diagnosis of a number of conditions, including leukemia, multiple myeloma, lymphoma, anemia, and pancytopenia.

- **Total protein test:** total protein test measures the amount of protein in your blood. Proteins are important for the health and growth of the body's cells and tissues. The test can help diagnose a number of health conditions.

- **C-reactive protein (CRP) test:** this is another test used to help diagnose conditions that cause inflammation.CRP is produced by the liver and if there is a higher concentration of CRP than usual, it's a sign of inflammation in your body.

- **Blood Electrolytes test:** An electrolyte panel is a blood test that measures the levels of electrolytes and carbon dioxide in blood, electrolytes are minerals found in the body, including sodium, potassium and chloride that perform jobs such as maintaining a healthy water balance in the body.

- **Urine albumin to creatinine ratio (ACR):** urine albumin to creatinine ratio (ACR), also known as urine microalbumin, helps identify kidney disease that can occur as a complication of diabetes.

- **Protein electrophoresis:** is used to identify the presence of abnormal proteins, to identify the absence of normal proteins, and to determine when different groups of proteins are present in unusually high or low amounts in blood or other body fluids. Protein electrophoresis separates proteins based on their size and electrical charge. This forms a characteristic pattern of bands of different widths and intensities on a test media and reflects the mixture of proteins present in the body fluid evaluated. The pattern is divided into five fractions, called albumin, alpha 1, alpha 2, beta, and gamma. In some cases, the beta fraction is further divided into beta 1 and beta 2.

- **Immunofixation electrophoresis (IFE):** The immunofixation blood test is used to identify proteins called immunoglobulins in blood. Too much of the same immunoglobulin is usually due to different types of blood cancer. Immunoglobulins are antibodies that help your body fight infection.

- **Bence Jones protein:** is a monoclonal globulin protein or immunoglobulin light chain found in the urine, with a molecular weight of 22-24 kDa. Detection of Bence Jones protein may be suggestive of multiple myeloma or Waldenström's macroglobulinemia.

- **Blood type test:** a blood sample is needed. The test to determine your blood group is called ABO typing.

- **Free light chain test:** this test can pick up small amounts of free light chains in the blood. Doctors measure the ratio of kappa light chains to lambda

light chains. If myeloma cells make kappa or lambda light chains, the level of that light chain is increased and the ratio becomes abnormal.

- **Serum calcium:** it is a blood test to measure the amount of calcium in the blood. Serum calcium is usually measured to screen for or monitor bone diseases or calcium-regulation disorders (diseases of the parathyroid gland or kidneys).

- **A serum creatinine test:** measures the level of creatinine in your blood and provides an estimate of how well your kidneys filter (glomerular filtration rate). If your kidneys aren't functioning properly, an increased level of creatinine may accumulate in your blood.

- **A blood urea nitrogen (BUN) test:** measures the amount of nitrogen in your blood that comes from the waste product urea. Urea is made when protein is broken down in your body. Urea is made in the liver and passed out of your body in the urine. A BUN test is done to see how well your kidneys are working.

- **Creatinine clearance test:** measures how well creatinine is removed from your blood by your kidneys. This test gives better information than a blood creatinine test on how well your kidneys are working. The test is done on both a blood sample and on a sample of urine collected over 24 hours.

- **Beta-2 microglobulin (B2M) test:** is used as a tumor marker for some people with blood cell cancers. It is not diagnostic for a specific disease, but it has been associated with the amount of cancer present (tumor burden) and can give a healthcare practitioner additional information about someone's likely prognosis. A blood B2M test and sometimes a urine test may be ordered to help determine the severity and spread (stage) of multiple myeloma, to help evaluate the prognosis of cancers such as multiple myeloma and lymphoma, and may sometimes be ordered to evaluate disease activity and the effectiveness of treatment. Recently, the International Myeloma Working Group published new guidelines called the International Staging System for Multiple Myeloma. The staging system is based mainly off of levels of both albumin and B2M in the blood. Higher blood B2M levels correspond with higher disease stages and therefore more advanced disease with worse prognosis.

- **Bilirubin Test:** bilirubin test measures how much bilirubin is in the blood. Bilirubin is made when red blood cells break down. The liver changes the bilirubin so that it can be excreted from the body. High bilirubin levels might mean there's a problem with the liver. In newborns, it can take some time for the liver to start working properly. High bilirubin levels can make skin and eyes look yellow, called jaundice.

- **Echo test:** An echocardiogram (echo) is a graphic outline of the heart's movement. During an echo test, ultrasound (high-frequency sound waves) from a hand-held wand placed on your chest provides pictures of the heart's

valves and chambers and helps the sonographer evaluate the pumping action of the heart.

- **ECG:** An electrocardiogram (ECG) is a medical test that detects cardiac (heart) abnormalities by measuring the electrical activity generated by the heart as it contracts. The machine that records the patient's ECG is called an electrocardiograph.
- **MRI Scan:** Magnetic resonance imaging (MRI) uses a large magnet and radio waves to look at organs and structures inside your body. Health care professionals use MRI scans to diagnose a variety of conditions, from torn ligaments to tumors. MRIs are very useful for examining the brain and spinal cord.
- **Radiography:** it is an imaging technique using X-rays, gamma rays, or similar radiation to view the internal form of an object.
- **PET scan:** Positron emission tomography (PET) scans are used to produce detailed 3-dimensional images of the inside of the body. The images can clearly show the part of the body being investigated, including any abnormal areas, and can highlight how well certain functions of the body are working.
- **CT scan:** Computerized tomography (CT) scans use X-rays and a computer to create detailed images of the inside of the body. CT scans are sometimes referred to as CAT scans or computed tomography scans.

2.3 Experiments and Results

The dataset contain 149 instances from three classes according to (ISS). The choice of matrix cost was quite subjective.We first perform our experiments on the MM dataset using the importance variables measures as features costs then we replace those importance variables measures with the real prices of each diagnostic test to see what can be changed.

Using the importance variables measures as features costs :

MULTIPLE MYELOMA		
Methods	**CostVimp**	**Random Forest**
Missclassification cost	0.03	0.08
Error rate	0.15	0.17
Selected variables	3	all
Budget	83.07%	/

Table 4.4: Results of Multiple myeloma using the importance variables measures as features costs

According to our model the three most important features are: Beta-2 microglobulin (B2M), Free light chain and Albumin.
Figure 4.1 show the final tree of CostVimp with only three selected variables Free light chain on top as split node, then Beta-2 microglobulin (B2M) and Albumin as its children nodes.

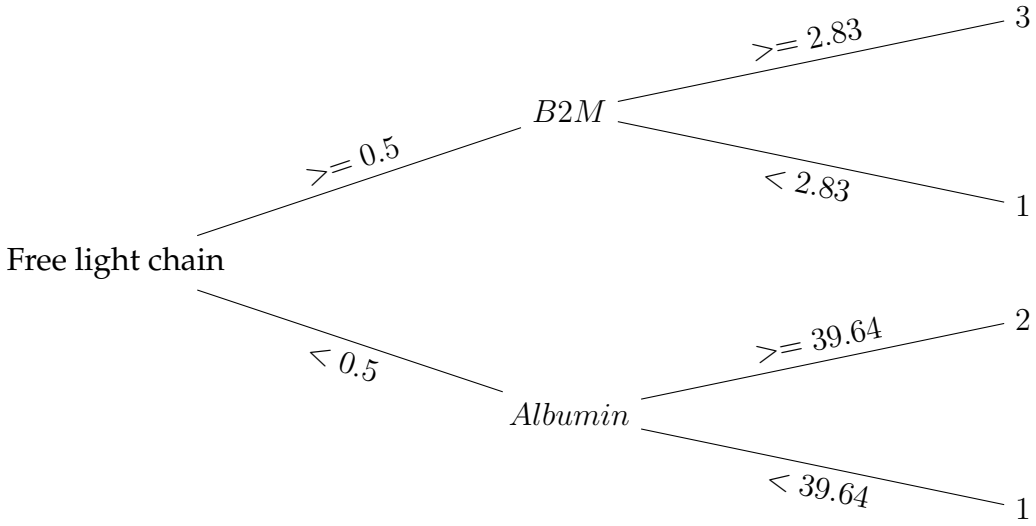

Figure 4.1: The optimal tree of CostVimp using the importance variables measures as features costs on MM dataset

Using the real prices:
The table below, 4.5 show the several used test to diagnostic the MM and its stage assigned to their real prices.

DIAGNOSTIC TEST	REAL PRICE
Complete Blood Count (CBC)	500 DA
Bone marrow examination	600 DA
Total protein test	300DA
C-reactive protein (CRP) test	600 DA
Blood Electrolytes test	900 DA
Protein electrophoresis	4100 DA
Immunofixation electrophoresis (IFE)	1600 DA
Bence Jones protein	1600 DA
Blood type test	300DA
Albumin	500DA

Free light chain test	1600 DA
Serum calcium	400 DA
A serum creatinine test	300 DA
A blood urea nitrogen (BUN) test	300 DA
Creatinine clearance test	300 DA
Beta-2 microglobulin (B2M) test	1200 DA
Bilirubin Test	2200 DA
Echo test	2000 DA
ECG	1500 DA
MRI Scan	18000 DA
Radiography	1500DA
CT scan	8000DA
Total	25300DA

Table 4.5: Diagnostic tests of MM & thier prices

Results of our experiments on MM dataset using the real prices are shown in 4.6.

MULTIPLE MYELOMA			
Error rate	**Misclassification cost**	**Budget**	**Selected Features**
0.21	0.02	1700/25300 DA	2
0.15	0.02	3300/25300 DA	3

Table 4.6: Results of Multiple myeloma using the real prices

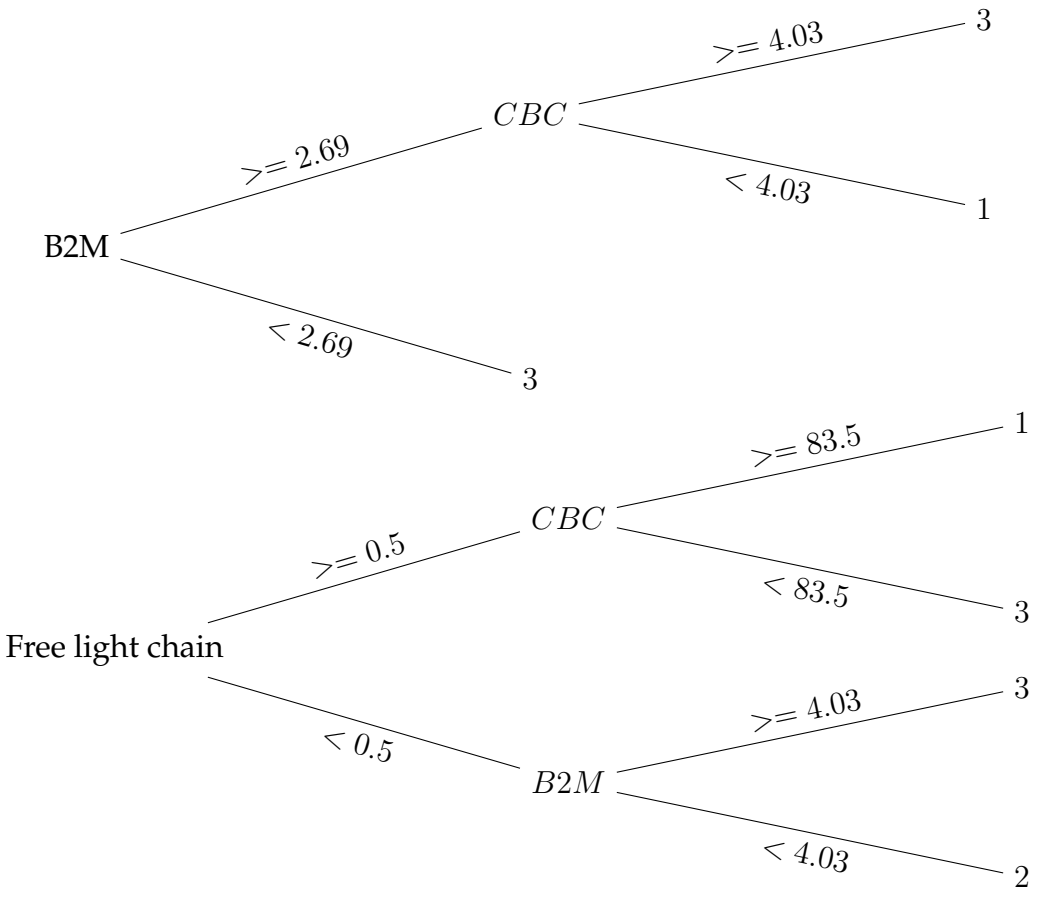

Figure 4.2: The optimal trees of CostVimp using the real prices.

As we can see in figure 4.2, both trees have selected Beta-2 microglobulin (B2M) and CBC tests to do the split, using the Free light chain test on the second tree (below), makes it more accurate. Unlike the tree above that knows only two classes, this experiment validate our previous note: the more variables we use the more our classifier is accurate.

The most expensive tests are not usually the most informative neither the most important ones and vice versa. From Experiment 1 and Experiment 2, we can say that the use of: Beta-2 microglobulin (B2M), CBC, free light chain and albumin is more than sufficient to distinguish MM stages, that means a total budget of 3800 DA instead of 25300 DA.

2.4 Comparaison results

In this section we are going to compare our model (Cost Vimp) with the Cost Sensitive Classification Tree Algorithm (CSCART) [60], Cost sensitive decision Tree Algorithm (CSC4.5) [63] on the MM real dataset using Tanagra platform [5]. Results are shown in Table 4.7

[5]https://eric.univ-lyon2.fr/ ricco/tanagra/fr/tanagra.html

Methods	**CSCART**			**CSC4.5**			**CostVimp**		
Missclassification cost	0.15			0.19			0.03		
Confusion matrix									
	Stage 3	Stage2	Strage1	Stage 3	Stage2	Strage1	Stage 3	Stage2	Strage1
Stage3	44	0	0	42	0	2	41	0	3
Stage2	4	0	0	3	0	1	4	0	0
Stage1	4	0	0	4	0	0	1	0	3

Table 4.7: Results of CostVimp Vs CSCART & CSC4.5

As show in Table 4.7 the Cost Sensitive Classification Tree (CSCART) Algorithm can recognize only one class (the most dominanted one in the dataset), while the CSC4.5 can recogize 2 classes (stage 1 and 3) but show a heigher missclassification cost. Comparing to our model the Cost Vimp algorithm outperform the two classifiers in terms of accuracy and cost.

3 Conclusion

In this chapter, we analyzed and discussed the performance of our model on ten databases: nine from UCI machine learning and one detailed study case from the real world: multiple myeloma dataset. Our model show promising results.

Conclusion

In machine learning, the field of cost-sensitive learning is recognized as an active domain of research that focuses on handling different types of costs. In the literature, a number of different approaches have been devised in order to deal with different types of costs, such as the cost of tests and misclassification costs. A number of academics have directed their efforts towards developing approaches and classifiers that consider misclassification costs; however, the most suitable cost-sensitive classifier for a given data set and problem remains unknown.

A number of the budgeted learning approaches and cost sensitive approaches have been devised and introduced during the last decade; however, establishing which are the most valuable is not simple, with no best method recognized amongst the options. Accordingly, this study has aimed to make a compromise between the cost of tests and misclassification cost by investigating the measure importance variable generated by the RF algorithm as test costs.

In the area of health, costs are direct or indirectly present in the majority of situations. A variety of financial or human costs can be associated with a specific diagnostic test. The utilization of learning methods for the generation of diagnostic or prognostic models, that are sensitive to several types of costs, is an important step to transform the computer based process of knowledge acquisition into a more natural process, with tendency to be similar with mental processes used by medical doctors. On the other hand, these kinds of strategies can permit large financial savings and benefits in health-related quality of life costs.

Hence this thesis has aimed to study the use of the measure importance variable in a cost sensitive algorithm to establish a link between the used variables (tests) and the total budget. As technologies became more expensive and budgets are limited, it is even more rational to consider all the cost involved. A big challenge is to have better healthcare using less money. Our proposed algorithm showed good results in terms of accuracy, time of execution, misclassification cost and total budget. A detailed case study on MM is given in the thesis.

This thesis focused on binary cost-sensitive classification problems. Nevertheless, not all cost-sensitive applications are two-class problems. Therefore, we expect that an interesting line of future work should extend to multi-class problems. It is also interesting to evaluate our strategies, with new experiments in other datasets, with real data and real costs.

Bibliography

[1] Kylie Urban, "Risk, Benefit or Cost: What Stops Patients from Receiving a Diagnostic Test?," https://labblog.uofmhealth.org/lab-report, June 09, 2017, [Online; accessed february-2019].

[2] Maxim Dmitrievsky, "RANDOM DECISION FOREST IN REINFORCEMENT LEARNING," https://www.mql5.com/en/articles/3856/, 6 July 2018, [Online; accessed may-2019].

[3] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[4] Peter D. Turney, "Types of cost in inductive concept learning," *CoRR*, vol. cs.LG/0212034, 2002.

[5] Daniel J. Lizotte, Omid Madani, and Russell Greiner, "Budgeted learning of naive-bayes classifiers," in *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, August 7-10 2003*, 2003, pp. 378–385.

[6] Bianca Zadrozny and Charles Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2001, KDD '01, pp. 204–213, ACM.

[7] Xiaoyong Chai, Lin Deng, Qiang Yang, and Charles X. Ling, "Test-cost sensitive naive bayes classification," in *Proceedings of the Fourth IEEE International Conference on Data Mining*, Washington, DC, USA, 2004, ICDM '04, pp. 51–58, IEEE Computer Society.

[8] Peter D. Turney, "Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm," *CoRR*, vol. cs.AI/9503102, 1995.

[9] Susan Lomax and Sunil Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Comput. Surv.*, vol. 45, no. 2, pp. 16:1–16:35, Mar. 2013.

[10] Pedro Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 1999, KDD '99, pp. 155–164, ACM.

[11] Kate McCarthy, Bibi Zabar, and Gary Weiss, "Does cost-sensitive learning beat sampling for classifying rare classes?," in *Proceedings of the 1st International Workshop on Utility-based Data Mining*, New York, NY, USA, 2005, UBDM '05, pp. 69–77, ACM.

[12] Kai Ming Ting and Zijian Zheng, "Boosting cost-sensitive trees," in *Proceedings of the First International Conference on Discovery Science*, London, UK, UK, 1998, DS '98, pp. 244–255, Springer-Verlag.

[13] Bianca Zadrozny, John Langford, and Naoki Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proceedings of the Third IEEE International Conference on Data Mining*, Washington, DC, USA, 2003, ICDM '03, pp. 435–, IEEE Computer Society.

[14] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[15] Nada Lavrac, Dragan Gamberger, and Peter D. Turney, "Cost-sensitive feature reduction applied to a hybrid genetic algorithm," in *Algorithmic Learning Theory, 7th International Workshop, ALT '96, Sydney, Australia, October 23-25, 1996, Proceedings*, 1996, pp. 127–134.

[16] Charles X. Ling, Qiang Yang, Jianning Wang, and Shichao Zhang, "Decision trees with minimal costs," in *ICML*. 2004, vol. 69 of *ACM International Conference Proceeding Series*, ACM.

[17] Shengli Sheng, Charles X. Ling, and Qiang Yang, "Simple test strategies for cost-sensitive decision trees," in *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, 2005, pp. 365–376.

[18] Shengli Sheng and Charles X. Ling, "Hybrid cost-sensitive decision tree," in *Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005, Proceedings*, 2005, pp. 274–284.

[19] Shengli Sheng, Charles X. Ling, Ailing Ni, and Shichao Zhang, "Cost-sensitive test strategies," in *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, 2006, pp. 482–487.

[20] Alberto Freitas, Altamiro da Costa Pereira, and Pavel Brazdil, "Cost-sensitive decision trees applied to medical data," in *DaWaK*. 2007, vol. 4654 of *Lecture Notes in Computer Science*, pp. 303–312, Springer.

[21] Valentina Bayer Zubek and Thomas G. Dietterich, "Pruning improves heuristic search for cost-sensitive learning," in *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, 2002, pp. 19–26.

[22] Andrew Arnt and Shlomo Zilberstein, "Attribute measurement policies for time and cost sensitive classification," in *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK*, 2004, pp. 323–326.

[23] Victor S. Sheng, "Fast data acquisition in cost-sensitive learning," in *Advances in Data Mining. Applications and Theoretical Aspects - 11th Industrial Conference, ICDM 2011, New York, NY, USA, August 30 - September 3, 2011. Proceedings*, 2011, pp. 66–77.

[24] Russell Greiner, Adam J. Grove, and Dan Roth, "Learning cost-sensitive active classifiers," *Artif. Intell.*, vol. 139, no. 2, pp. 137–174, 2002.

[25] Johannes Leveling, Giorgio Maria Di Nunzio, and Thomas Mandl, "Logclef: enabling research on multilingual log files," in *Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval, PMHR@HT 2011, Eindhoven, The Netherlands, June 6, 2011*, 2011, pp. 55–56.

[26] Yanmin Sun, Mohamed S. Kamel, Andrew K. C. Wong, and Yang Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[27] Gary M. Weiss, Kate McCarthy, and Bibi Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?," in *Proceedings of the 2007 International Conference on Data Mining, DMIN 2007, June 25-28, 2007, Las Vegas, Nevada, USA*, 2007, pp. 35–41.

[28] Lev Reyzin, "Boosting on a budget: Sampling for feature-efficient prediction," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 2011, pp. 529–536.

[29] He He and Hal Daumi, "Cost-sensitive dynamic feature selection," 2012.

[30] Zhixiang Eddie Xu, Kilian Q. Weinberger, and Olivier Chapelle, "The greedy miser: Learning under test-time budgets," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[31] Sergey Karayev, Mario J. Fritz, and Trevor Darrell, "Dynamic feature selection for classification on a budget," 2013.

[32] Ma Li and Suohai Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, pp. 169:1–169:18, 2017.

[33] Shihao Ji and Lawrence Carin, "Cost-sensitive feature acquisition and classification," *Pattern Recognition*, vol. 40, no. 5, pp. 1474–1485, 2007.

[34] Shuang Wu, Xiaofeng Gao, and Guihai Chen, "Some transformation methods on probabilistic model for crowdsensing networks," in *Fifth IEEE International Conference on Big Data and Cloud Computing, BDCloud 2015, Dalian, China, August 26-28, 2015*, 2015, pp. 304–309.

[35] Sergey Karayev, Tobias Baumgartner, Mario Fritz, and Trevor Darrell, "Timely object recognition," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 899–907.

[36] Feng Nan, Joseph Wang, and Venkatesh Saligrama, "Feature-budgeted random forest," in *ICML*. 2015, vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 1983–1991, JMLR.org.

[37] Feng Nan and Venkatesh Saligrama, "Adaptive classification for prediction under a budget," in *NIPS*, 2017, pp. 4730–4740.

[38] Gabriella Contardo, *Machine learning under budget constraints. (Apprentissage statistique sous contraintes de budget)*, Ph.D. thesis, Pierre and Marie Curie University, Paris, France, 2017.

[39] Mohammad Kachuee, Orpaz Goldstein, Kimmo Karkkainen, Sajad Darabi, and Majid Sarrafzadeh, "Opportunistic learning: Budgeted cost-sensitive learning from data streams," *CoRR*, vol. abs/1901.00243, 2019.

[40] Prem Melville, Maytal Saar-Tsechansky, Foster J. Provost, and Raymond J. Mooney, "Active feature-value acquisition for classifier induction," in *ICDM*. 2004, pp. 483–486, IEEE Computer Society.

[41] Prem Melville, Saharon Rosset, and Richard D. Lawrence, "Customer targeting models using actively-selected web content," in *KDD*. 2008, pp. 946–953, ACM.

[42] Maytal Saar-Tsechansky, Prem Melville, and Foster J. Provost, "Active feature-value acquisition," *Management Science*, vol. 55, no. 4, pp. 664–684, 2009.

[43] Kun Deng, Joelle Pineau, and Susan A. Murphy, "Active learning for developing personalized treatment," *CoRR*, vol. abs/1202.3714, 2012.

[44] Aloak Kapoor and Russell Greiner, "Reinforcement learning for active model selection," in *Proceedings of the 1st International Workshop on Utility-based Data Mining*, New York, NY, USA, 2005, UBDM '05, pp. 17–23, ACM.

[45] Sudipto Guha and Kamesh Munagala, "Approximation algorithms for budgeted learning problems," in *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, 2007, pp. 104–113.

[46] Gianluca Bontempi and Olivier Caelen, "A selecting-the-best method for budgeted model selection," in *ECML/PKDD (1)*. 2011, vol. 6911 of *Lecture Notes in Computer Science*, pp. 249–262, Springer.

[47] Haiqin Yang, Zenglin Xu, Michael R. Lyu, and Irwin King, "Budget constrained non-monotonic feature selection," *Neural Networks*, vol. 71, pp. 214–224, 2015.

[48] Feng Nan, Joseph Wang, and Venkatesh Saligrama, "Feature-budgeted random forest," *CoRR*, vol. abs/1502.05925, 2015.

[49] Besmira Nushi, Adish Singla, Andreas Krause, and Donald Kossmann, "Learning and feature selection under budget constraints in crowdsourcing," in *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA.*, 2016, pp. 159–168.

[50] Feng Nan, Joseph Wang, and Venkatesh Saligrama, "Pruning random forests for prediction on a budget," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 2334–2342.

[51] Hajin Shim, Sung Ju Hwang, and Eunho Yang, "Joint active feature acquisition and classification with variable-size set encoding," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 2018, pp. 1375–1385.

[52] Leo Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[53] Leo Breiman, "Pasting small votes for classification in large databases and on-line," *Machine Learning*, vol. 36, no. 1-2, pp. 85–103, 1999.

[54] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] Zhi-Hua Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall/CRC, 1st edition, 2012.

[56] Thomas G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, London, UK, UK, 2000, MCS '00, pp. 1–15, Springer-Verlag.

[57] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[58] Lior Rokach and Oded Maimon, *Data Mining with Decision Trees: Theroy and Applications*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008.

[59] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2 edition, 2009.

[60] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA, 1984, new edition [**?**]?

[61] Lior Rokach and Oded Maimon, "Classification trees," in *Data Mining and Knowledge Discovery Handbook*, pp. 149–174. Springer, 2010.

[62] J. R. Quinlan, "Learning with continuous classes," 1992, pp. 343–348, World Scientific.

[63] Jeffrey P. Bradford, Clayton Kunz, Ron Kohavi, Clifford Brunk, and Carla E. Brodley, "Pruning decision trees with misclassification costs," in *Proceedings of the 10th European Conference on Machine Learning*, London, UK, UK, 1998, ECML '98, pp. 131–136, Springer-Verlag.

Résumé

De nos jours, avoir de bons soins de santé en utilisant moins d'argent devient un défi, car les technologies sont devenues de plus en plus coûteuses et les budgets sont limités. D'autre part, dans le diagnostic médical, une fausse prédiction négative (une personne malade déclarée comme étant saine) peut avoir des conséquences plus graves qu'une fausse prédiction positive et leur attribuer des coûts égaux est inexacte.
Ce projet de fin d'études contribue à la fois aux apprentissages budgétisés et aux apprentissages sensibles au coût en développant un modèle capable de faire un compromis entre les coûts de classification erronée et les coûts de test. Le modèle proposé est basé sur l'idée d'utiliser les mesures d'importance de variables de la forêt aléatoire en tant que coûts de test et en choisissant l'arbre optimal de la forêt développée en tant que stratégie de test. Notre modèle a été testé sur dix base de données : neuf bases de données de UCI Machine Learning et une base de données du monde réel : le myélome multiple ; collectée au Centre de Lutte Contre le Cancer (CLCC) de Tlemcen.
Mots clés :

Apprentissage sensible au coût, apprentissage budgétisé, forêts aléatoires, mesures d'importance des variables, UCI Machine Learning, Myélome multiple.

Abstract

Nowdays, having a good healthcare using less money become a challenge, as technologies became more and more expensive and budgets are limited. On the other hand, in the medical diagnosis, a false negative prediction (a sick person declared as healthy one) may have more serious consequences than a false positive prediction and assigning them equal costs is probably incorrect.
This Master thesis makes contributions to both the fields of budgeted-learning and cost sensitive learning in that it develops a model that can make a compromise between misclassification costs and test costs at the same time.
The proposed model is based on the idea of using the variables importance measures of random forest as test costs and choosing the optimal tree from the grown forest as test strategy. Our model has been tested on nine UCI Machine Learning datasets and on a real-world database: multiple myeloma; collected from the anti-cancer center of Tlemcen.

Keywords
Cost sensitive learning, budgeted learning, random forests, variables importance measures, UCI Machine Learning Datasets, Multiple myeloma.

الملخص

في الوقت الحاضر اصبح الحصول على رعاية صحية جيدة باستخدام أموال اقل يمثل تحديا، حيث أصبحت التقنيات باهظة الثمن بشكل متزايد و الميزانيات محدودة.
من ناحية أخرى في تشخيص طبي قد يكون للتنبؤ السلبي الخاطئ (شخص مريض يشخص أنه يتمتع بصحة جيدة)عواقب وخيمة اكثر من التنبؤ الإيجابي الخاطئ، ومن المحتمل أن يكون اعتبار تكاليف هذين الاثنين متساوية، غير مضبوط.
تساهم هذه الأطروحة في كل من التعلم المدرج في الميزانية وتعلم الحساس من حيث التكلفة من خلال تطوير نموذج يمكن أن يحدث المفاضلة بين تكاليف تصنيف الخاطئ والتكاليف الاختبار يعتمد النموذج المقترح على فكرة استخدام مقاييس أهمية متغير غابات العشوائية كالتكاليف الاختبار و اختيار الشجرة المثلى للغابات المطورة كاستراتيجية اختبار.

تم اختبار نموذجنا على 10 قواعد بيانات :تسع قواعد بيانية من UCI machine learning وقاعدة بيانية في العالم الحقيقي:
الورم النخاعي المتعدد التي تم جمعها في مركز مكافحة سرطان بتلمسان.

الكلمات المفتاحية:

التعلم حساس من حيث التكلفة التعلم في الميزانية الغابة العشوائية مقاييس أهمية متغيرات الورم النخاعي المتعدد