Apprentissage statistique de données relationnelles

Mickael Poussevin

Rédacteur	Mickaël Poussevin
Ecole	TELECOM Bretagne
Formation	Ingénieur Filière SLR et Master 2 Recherche, Systèmes informatiques centrés sur l'humain.
Date	13 août 2010

Rapport de stage de fin d'études : Apprentissage statistique de données relationnelles

Maître de stage :	Patrick Gallinari
Lieu du stage :	LIP6
Début du stage :	29 mars 2010
Fin du stage :	30 septembre 2010
Tuteur enseignant :	Antoine Beugnard

REMERCIEMENTS

Ayant été confronté pour la première fois à une tâche de recherche à l'issue d'une formation d'ingénieur, je tiens à remercier les personnes suivantes, qui m'ont grandement aidé à franchir cette étape et motivé à continuer en thèse :

- Patrick Gallinari, pour m'avoir ouvert les portes du laboratoire, conseillé et écouté pendant tout le déroulement du stage.
- Ludovic Denoyer, chercheur de l'équipe MALIRE, pour avoir pris du temps pour discuter avec moi et développer le modèle de la classification par paquet.
- Antoine Beugnard pour ses réponses à mes questions tout au long du stage et plus généralement de mon année de Master.
- Les stagiaires et doctorants de l'équipe MALIRE avec lesquels j'ai travaillé, et en particulier Gabriel Arnold Dulac et Amine Mekki.

RESUME

Ce document résume l'expérience que j'ai vécue lors du stage de fin d'études que j'ai réalisé de fin mars à septembre au Laboratoire d'Informatiques de Paris 6. Dans l'organisation de ce vaste laboratoire, l'équipe MALIRE du département DAPA est spécialisée dans le domaine de l'apprentissage automatique et c'est dans ce contexte que j'ai évolué.

Le travail que j'ai accompli est une tâche exploratoire autour du sujet des données relationnelles dynamiques qui semblent être une prochaine étape pour le domaine de l'apprentissage artificiel. La première partie en est une étude bibliographique qui me permit de réaliser que pour l'instant il n'existe pas de cadre clair pour ce type de données et de formuler une problématique générale : comment appréhender les relations entre données et introduire le temps dans la classification ? Elle fut suivie par le développement et la formalisation d'un modèle, celui de la classification par paquets, de la formalisation mathématique à la réalisation d'expériences pour le tester.

Je présente ensuite une analyse d'un problème du domaine : celui de la disponibilité des données qui pose la question des relations entre entreprises et laboratoires de recherche autour de l'apprentissage artificiel.

Les difficultés rencontrées au cours de se stage que ce soit autour de l'organisation et la rigueur dans le travail ou dans la recherche bibliographique ou de données, m'ont permis de confronter ma formation plutôt orientée ingénieur à la réalisation d'une tâche de recherche et m'ont préparés à poursuivre cette expérience au cours de ma thèse.

ABSTRACT

This document gathers the experience I lived as an intern in the Laboratoire d'Informatique de Paris 6. This laboratory is divided in five departments, each of them concerned by a different field in computer science. As my domain of interest was machine learning, I joined the MALIRE team, part of the DAPA department for this six months' internship, starting in March.

I accomplished there an exploratory task around relational temporal data, which seems to be one of the next steps of the field. A bibliographical study revealed that no framework or model was clearly established for machine learning on relational temporal data. Thus, I could formulate a general problematic: how to take advantage of links between pieces of data and introduce time in classification? The development from scratch of a model, named packed classification, tested by experiments, tried to answer this problematic.

I also analyzed a problem I was confronted with during this internship: gathering data. This problem seems recurrent in machine learning and raises the question of the links between data-rich company and laboratories.

Many difficulties crossed my path during this internship, of which the necessity of organization and rigor or the gathering knowledge or data are two representative examples, forced me to confront my engineering course to the realization of a research task and prepared me to carry on this experience during my PhD.

Sommaire

Table des figures

1. INTRODUCTION

Le stage de six mois que j'ai effectué au sein du Laboratoire d'Informatique de Paris 6, sous la responsabilité de Patrick Gallinari (LIP6) et Antoine Beugnard (TELECOM Bretagne) et que je rapporte dans ce document, conclut la formation d'ingénieur chercheur que j'ai suivie à TELECOM Bretagne. Il s'agit d'une activité de recherche dans le domaine de l'apprentissage artificiel autour de la problématique des données relationnelles et de la prise en compte de leur dynamique.

Ce rapport vise à présenter dans un premier temps le contexte, tant au point de vue du laboratoire que du domaine, au sein duquel j'ai évolué en répondant à deux questions : quelle est l'organisation du LIP6 ? Qu'est-ce que l'apprentissage artificiel aujourd'hui ? Puis il présente le résultat du travail que j'ai réalisé : une étude bibliographique qui me permit de dégager une problématique générale suivie de la formalisation d'un modèle permettant d'y répondre et des expériences permettant de vérifier son fondement. Il se termine par une analyse de la difficulté à trouver des corpus de données.

2. CONTEXTE

2.1 LABORATOIRE D'INFORMATIQUE DE PARIS 6

J'ai effectué mon stage au sein du LIP6 (Laboratoire d'Informatique de Paris 6), sous la responsabilité de Patrick Gallinari, actuellement professeur à l'UPMC (Université Pierre et Marie Curie) et directeur du laboratoire.

2.1.1 Présentation du laboratoire

Le LIP6 est considéré comme l'un des plus importants laboratoires de recherche en informatique. Il regroupe 185 chercheurs permanents et 256 doctorants sous la double tutelle UPMC et CNRS (UMR 7606). Sa taille est à la fois un atout pour augmenter sa visibilité à l'international et un possible inconvénient car une organisation claire et efficace est nécessaire afin de fournir un bon environnement de travail aux chercheurs.

Organigramme

L'organigramme présenté Figure 1 résume l'organisation du laboratoire. Il est composé de cinq départements qui recouvrent une large partie des thématiques de recherche actuelles, d'une équipe administrative de 13 personnes et d'une équipe technique de 12 employés.

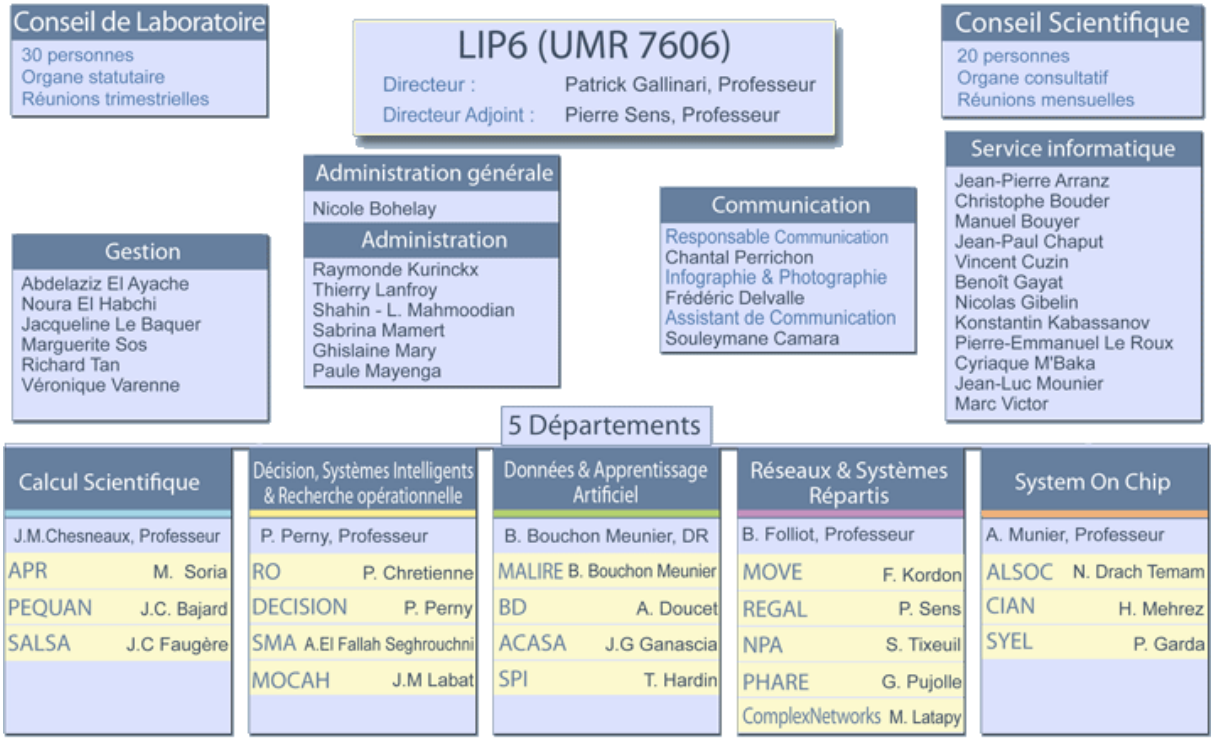

Figure 1 : organigramme du LIP6.

Equipes administratives et techniques

Ces deux équipes ont un rôle important dans le bon fonctionnement du laboratoire. La première s'occupe de la gestion des ressources humaines. La gestion du personnel est complexe dans le cas d'un établissement public avec un grand nombre de dossier à compléter et à envoyer à la bonne personne (inscriptions en thèse, missions pour assister à des conférences, organisation de conférences, …). Il faut de plus gérer les avancements au niveau UPMC et/ou CNRS suivant l'affiliation des chercheurs. La seconde surveille l'état du parc informatique du laboratoire, ce qui est une tâche critique dans le cas d'un établissement de recherche en informatique.

Département Calcul Scientifique

Ce département (LIP6, 2010) regroupe deux équipes SPIRAL (Systèmes Polynomiaux, Implantation, Résolution ALgébrique) et PEQUAN (PErformance et QUalité d'Algorithmes Numériques) qui traitent respectivement des problèmes de calcul symbolique et numérique et l'équipe SALSA et est sous la responsabilité de Jean-Marie Chesnaux.

Département DEcision, Systèmes Intelligents Recherche opérationnelle (DESIR)

Organisé en cinq équipes, RO (Recherche Opérationnelle), DECISION, SMA (Systèmes Multi-Agents), AnimatLab et MOCAH (Modèles et Outils en ingénierie des Connaissances pour l'Apprentissage Humain), ce département (LIP6, 2010) sous la responsabilité de Patrice Perny, traite des problématiques de décision, d'optimisation et de systèmes adaptatifs.

Département Données et APprentissage Artificiel (DAPA)

Comme son nom l'indique, le département (LIP6, 2010) étudie les questions d'apprentissage automatique, qu'il soit statistique ou symbolique, et les différentes manières de stocker l'information, en particulier de façon distribuée dans un système pair à pair, et est divisé en quatre équipes : MALIRE (MAchine Learning and Information REtrieval), BD (Bases de Données), ACASA (Agents Cognitifs et Apprentissage Symbolique Automatique) et SPI (Systèmes Preuves Implémentations). Il est dirigé par Bernadette Bouchon-Meunier.

Département Réseaux et Systèmes Répartis

Sous la responsabilité de Bertil Folliot, se concentre, selon ce dernier (LIP6, 2010), « sur la conception de solutions pour construire et gérer les réseaux et systèmes du futur ». Il comprend cinq équipes MoVe (Modélisation et Vérification), REGAL (Répartition et Gestion des Applications à Large Echelle), NPA (Networks and Performances Analysis), Phare et ComplexNetworks.

Département Systèmes Embargués sur Puce

Le département (LIP6, 2010) développe méthodes et outils pour les systèmes multiprocesseurs intégrés sur puce. L'équipe ALSOC (Architecture et Logiciels pour Systèmes Embarqués sur Puce) s'occupe du niveau système et l'équipe CIAN (Circuits Intégrés Numériques et Analogiques) du niveau circuit.

2.1.2 Département DAPA, équipe MALIRE

Le domaine de mon stage était l'apprentissage statistique de données relationnelles. J'ai donc rejoint les stagiaires de l'équipe MALIRE, membre du département DAPA comme présenté dans la section précédente.

Les chercheurs, doctorants et stagiaires des différentes équipes du laboratoire sont regroupés dans des salles proches. Cela à pour effet de créer une communauté intéressée par des thématiques proches. Ces communautés entrainent des interactions entre les personnes et permettent au niveau des individus de pouvoir confronter leurs idées ou trouver de l'aide, au niveau des équipes d'avoir une vision cohérente de la recherche qu'elles produisent. Au sein de l'équipe MALIRE, j'ai notamment travaillé aux côtés de stagiaires intéressés par la diffusion d'informations et la sélection de caractéristiques. Mais cela entraine inévitablement un manque de communication entre équipes au sein du laboratoire, ce qui peut être problématique dans le cas de thématiques larges, comme les réseaux dynamiques, qui peuvent intéresser plusieurs équipes dans des départements différents. Cela demande donc une communication interne importante que ce soit par une lettre d'information interne, qui peut malgré tout ne pas être lue, ou mieux par des journées thématiques, comme celle organisée autour des réseaux dynamiques, qui permettent de rassembler des personnes et créer des liens entre équipes différentes.

Au sein de l'équipe MALIRE, j'ai été essentiellement en contact avec des personnes intéressées par la problématique de l'apprentissage dans ou proches des graphes. Damien Fouquet travailla sur la diffusion d'information, notamment dans des réseaux sociaux comme Twitter. Gabriel Arnold-Du Lac s'occupait de la sélection de caractéristiques, qui est une problématique générale et donc applicable aux données relationnelles mais aussi intéressante si l'on considère des données dynamiques dont les caractéristiques peuvent évoluer dans le temps.

2.2 APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique, *machine learning* en anglais, est un domaine de l'intelligence artificielle que j'ai découvert au cours du cursus de Master Systèmes Informatiques Centrés sur l'Humain.

2.2.1 Présentation

Pour présenter le domaine, je commencerai par une citation (Mitchell, 1997) de Tom M. Mitchell, actuellement responsable du département d'apprentissage automatique de *Carnegie Mellon University*. Il s'agit d'améliorer la performance $\mathcal{P}$ d'une machine à réaliser la tâche $\mathcal{T}$ en utilisant un ensemble d'exercices $\mathcal{E}$.

Les multiples façons d'exprimer ce triplet $(\mathcal{P}, \mathcal{T}, \mathcal{E})$ permettent de développer de nombreux cadres théoriques pour le domaine, dont celui de l'apprentissage statistique, de développer différents modèles par exemples génératifs ou discriminants, de les appliquer à diverses tâches comme les échecs, la conduite automobile et la classification.

Classification

Cette dernière tâche est une des plus étudiée dans le domaine. Son concept est à la fois simple, il s'agit de donner la bonne étiquette ou classe à une donnée en fonction de ses caractéristiques, et très général, beaucoup de tâches peuvent s'y réduire.

Pour définir un problème de classification, on commence par représenter les données dans l'espace vectoriel de leurs caractéristiques. On peut alors réaliser des calculs sur cette représentation des données et en particulier leur attribuer un score qui pourra ensuite être traduit en classe.

Ensuite, on identifie le nombre de classes présentes, si cela est possible. S'il y en a deux, on parle de classification binaire. S'il y en a plus, on parle de classification multi-classe et le problème est en général ramené à de la classification binaire. Par exemple en un contre tous, on définit autant de problèmes que de classes, chacun comparant l'appartenance à une classe face au reste. On assigne alors la classe qui aura été choisie avec le plus de confiance.

La généricité de la tâche de classification vient de la largesse du concept de classe qui peut revêtir de multiples formes. Il peut s'agir du sujet d'un document, d'actions de conduite, de coups dans un jeu. Il permet donc de ramener nombre de tâches à de la classification.

Apprentissages

On distingue trois grandes familles d'apprentissage, qui sont autant de façons d'exprimer $\mathcal{E}$ dans le triplet énoncé plus haut et qui s'adaptent à des cas réels différents.

La première forme est l'apprentissage supervisé. On dispose d'un oracle qui définit les classes de toutes les données de l'ensemble d'entrainement en fonction de leurs caractéristiques. Le but ici est d'apprendre à recréer le raisonnement de l'oracle afin d'être ensuite capable de l'appliquer sur d'autre données. Bien sûr, il n'est en général possible que d'approximer le raisonnement de l'oracle et la distance entre le raisonnement et son approximation s'appelle le biais.

La deuxième forme est l'apprentissage non-supervisé. Par opposition au précédent, on ne dispose pas d'oracle. Il s'agit alors de trouver des structures à l'intérieur des données. Une grande sous famille de l'apprentissage non-supervisé est le *clustering*. Il s'agit de trouver une partition des données, *cluster*, en général en définissant une notion de proximité entre les données. On peut utiliser des algorithmes comme ceux du ou des plus proche(s) voisin(s).

La troisième forme est l'apprentissage semi-supervisé. Dans ce cadre, seul une partie des étiquettes de l'ensemble d'entrainement sont connues. Il s'agit d'une forme de plus en plus utilisée car de nombreuses données réelles ne sont pas étiquetées et le travail d'oracle est beaucoup trop long et coûteux pour être réalisé entièrement.

Performances

La mesure des performances révèle le but que doit atteindre la machine. Prenons le cas d'un programme apprenant à jouer aux échecs. On peut penser dans un premier temps à deux mesures de performances : le nombre de parties gagnées sur le nombre de parties jouées ou le nombre de coup jugés très bon sur le nombre de coups joués. Chacune introduit un biais différent. La première cherche la victoire à tout prix, mais ne prend pas en compte le niveau de l'opposant. On peut se demander en effet si une victoire face à un débutant et une défaite face à un expert ont la même valeur. La seconde s'intéresse à produire le plus possible de très bons coups, sans s'intéresser à la victoire et s'appuie sur l'hypothèse que joue de très bons coups permet de bien jouer, qui n'est à priori pas certaine et demande en plus de savoir les reconnaître.

Dans le cadre de la classification la première mesure de performance est la mesure naïve du nombre de documents bien classés sur le nombre de documents totaux.

Cette mesure est générale est donne une première vue de l'efficacité de l'algorithme utilisé. Par contre, elle manque de précision et ne permet pas de savoir s'il y a un déséquilibre entre les classes : par exemple un classe peu nombreuse qui ne serait jamais reconnue n'affecterait pas beaucoup cette mesure.

On choisit en général d'utiliser les mesures de précision, rappel et F-score (Wikimedia, 2010).

Précision

On suppose que le problème comporte n classes et on définit la précision d'un algorithme de classification en fonction d'une classe i :

$$Precision_i = \frac{nombre\ de\ documents\ correctement\ attribues\ a\ la\ classe\ i}{nombre\ de\ documents\ attribues\ a\ la\ classe\ i}$$

La précision totale est définie comme suit :

$$Precision = \frac{\sum_{i=1}^{n} Precision_i}{n}$$

Rappel

On se place dans le même cadre et on définit le rappel pour une classe i :

$$Rappel_i = \frac{nombre\ de\ documents\ correctement\ attribues\ a\ la\ classe\ i}{nombre\ de\ documents\ appartenants\ a\ la\ classe\ i}$$

Puis le rappel total :

$$Rappel = \frac{\sum_{i=1}^{n} Rappel_i}{n}$$

F-score

On définit alors un résumé de ces deux valeurs, le F-score noté F_β, avec $\beta \in]0, +\infty[$.

$$F_\beta = (1 + \beta^2).\frac{Precision.Rappel}{\beta^2.Precision + Rappel}$$

Le paramètre β permet d'amplifier l'importance de l'un ou de l'autre des paramètres. Lors que les deux sont d'équivalente importance, on choisit $\beta = 1$.

Cette mesure, qui peut être utilisée dans le contexte des moteurs de recherche, permet de quantifier la capacité à correctement étiqueter chaque classe, avec le rappel, et la netteté des frontières entre les classes.

2.2.2 Données vectorielles

A ces débuts, l'apprentissage statistique considérait des données vectorielles, chacune étant représentée par son vecteur de caractéristiques. La majorité des méthodes peuvent alors être classées en deux catégories, génératives ou discriminantes. Pour présenter ces familles, on se place dans le cadre de l'apprentissage statistique supervisé.

Méthodes génératives

On peut se représenter ces méthodes comme essayant de définir des règles de raisonnements pour reproduire celui de l'oracle. D'un point de vue formel il s'agit d'estimer la probabilité conditionnelle $p(y|x)$ d'avoir la classe y en étant la donnée x, qui se révèle en pratique difficile à apprendre.

Théorème de Bayes

Ce théorème permet d'exprimer une probabilité conditionnelle $P(A|B)$ en fonction de $P(B|A)$. On en présente ici une formulation générale, pour une partition (A_i) des événements possibles et une variable aléatoire B, on a :

$$\forall i, \qquad P(A_i|B) = \frac{P(B|A_i).P(A_i)}{\sum_j \left(P\left(B|A_j\right).P\left(A_j\right)\right)}$$

Remarque : le dénominateur peut-être vu comme une constante.

Algorithmes Naïfs Bayésiens

Ces algorithmes utilisent le théorème présenté plus haut et apprennent les quantités $p(y)$, la probabilité d'avoir la classe y, et $p(x_i|y)$ la probabilité d'avoir la caractéristique x_i (on note

$x = (x_i)$) en étant de la classe y. Ils réalisent alors l'hypothèse naïve, d'où ils tirent leur nom, de considérer les différentes caractéristiques des données indépendantes entre-elles.

La formulation de leur problème d'inférence revient à chercher l'étiquette la plus probable en utilisant le théorème de Bayes, soit, pour une donnée x :

$$argmax_y \left(P(Y = y). \prod_i P(x_i | Y = y) \right)$$

Cette hypothèse est généralement fausse, si l'on considère par exemple les données d'un réseau de capteurs mesurant des informations météos, comme par exemple dans le laboratoire d'Intel (Peter Bodik, 2004), qui sont interdépendantes. Pourtant ils fournissent de bons résultats. Cela vient du fait que les relations entre les caractéristiques sont une information supplémentaire aux valeurs de ces caractéristiques, et exprimer mathématiquement ces relations est difficile. On peut donc choisir de s'en affranchir et de se priver du gain d'information, ce qui donne des résultats corrects mais aussi permet de ne pas tomber dans le piège d'une mauvaise expression des relations qui compromettrait fortement les performances.

Méthodes discriminantes

A l'inverse des méthodes génératives qui essaient de créer des règles de raisonnement pour en déduire des limites dans l'espace des données, les méthodes discriminantes apprennent directement les limites que fixe l'oracle. En général il s'agit d'hyperplan dans l'espace des données.

D'un point de vue formel, ces méthodes cherchent à apprendre directement $p(y|x) \geq 0{,}5$.

Séparateurs à Vaste Marge

Les algorithmes les plus classiques de méthodes discriminantes sont les SVMs (séparateurs à vaste marge), encore appelées machines à vecteurs de support. Ces deux noms en font ressortir le fonctionnement :

- trouver des vecteurs dits de support, qui sont les plus proches de la frontière ;
- trouver un hyperplan séparateur qui maximise la marge laissé entre lui et les vecteurs de support.

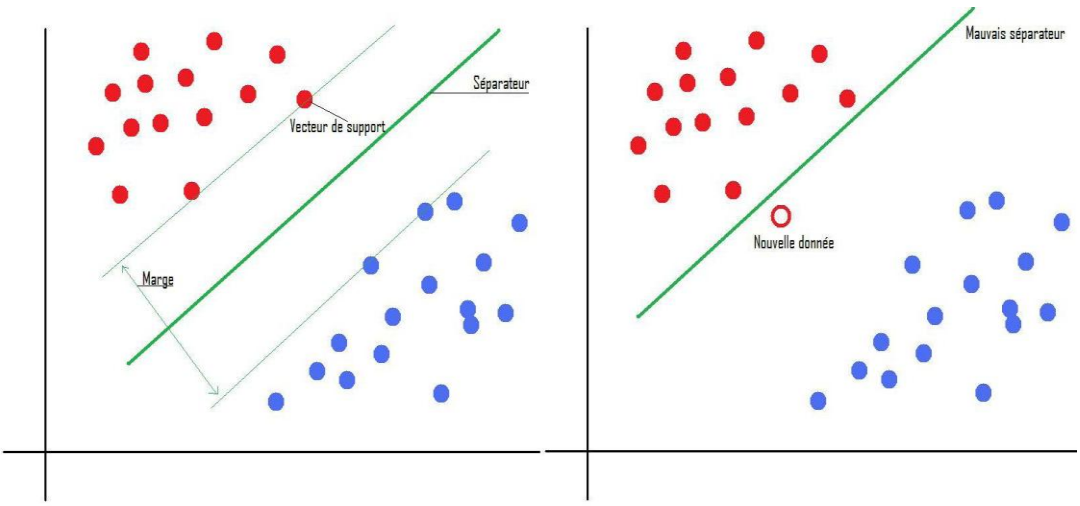

Figure 2 : présentation des machines à vaste marge

Comme on le voit Figure 2, cela permet d'éviter ensuite des erreurs lors de l'inférence, qui consiste à savoir dans quelle zone se trouve la nouvelle donnée.

Noyau

Comme on peut le constater, dans la Figure 2, les données sont linéairement séparables, c'est-à-dire qu'il existe un hyperplan permettant de séparer l'espace en deux zones, chacune contenant un seul type de données. Ce n'est pas toujours possible, et c'est en général dû au fait que l'espace des données est de petite dimension. Une solution consiste alors à les plonger dans un espace de plus grande dimension.

Ce que l'on appelle le noyau, que l'on note en général K, est le produit scalaire dans ce nouvel espace. On peut montrer qu'il est en fait possible d'exprimer directement le problème d'inférence avec le noyau, sans connaître la fonction utilisée pour plonger les données dans un espace plus grand.

Pour cela, il est nécessaire d'exprimer dans un premier temps le problème d'optimisation correspondant à l'inférence, puis d'utiliser le théorème de Karush-Kuhn-Tucker pour le reformuler sous la forme :

$$\begin{cases} \max_{\alpha} \left(\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right) \\ \forall i, \forall \alpha_i \geq 0, \qquad \sum_i \alpha_i y_i = 0 \end{cases}$$

2.2.3 Données relationnelles

Dans de nombreux problèmes réels, les données présentent des liens les une entre les autres. Ces liens sont généralement porteurs d'information. Les méthodes vectorielles choisissent de ne pas considérer ces relations et n'exploitent donc pas cette information. Les méthodes d'apprentissage statistiques sur des données relationnelles ont été développées pour les prendre en compte de manière pertinente.

Hypothèse de consistance

Ces méthodes considèrent en général l'ensemble des données représentées sous forme de graphe. Les nœuds sont les données, les arêtes, les liens entre ces dernières. Elles s'appuient sur une hypothèse, dite de consistance, qui prend deux formes :

- les données voisines dans un graphe sont susceptibles d'avoir une étiquette proche ;
- les données appartenant à une même sous-structure (comme une partie connexe, une clique,…) sont susceptibles d'avoir une étiquette proche.

Le problème des deux lunes, que l'on retrouve sur la couverture du livre d'Olivier Chapelle, ancien doctorant au Laboratoire (Chapelle, 2006), illustre cette hypothèse : chaque lune représentant une sous-structure.

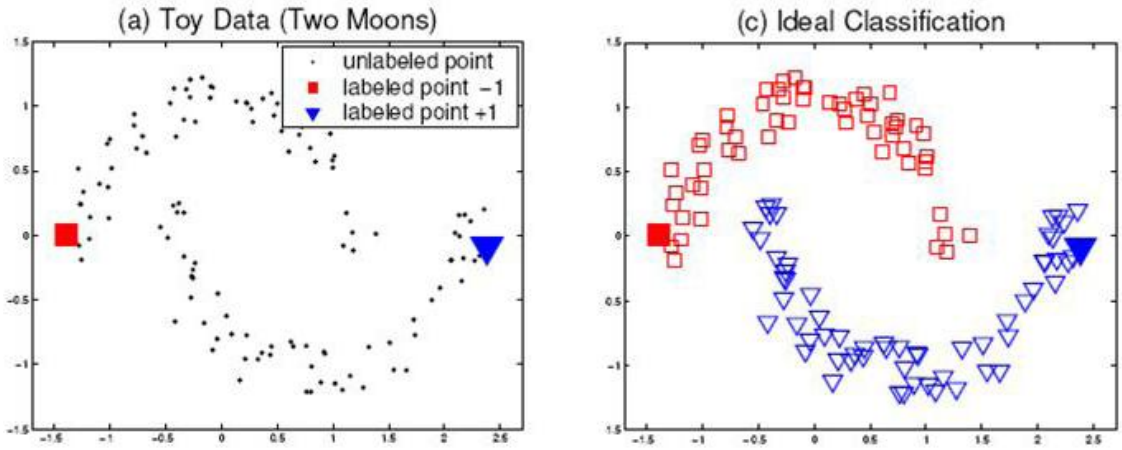

Figure 3 : le problème des deux lunes, extrait de (Zhou, 2004)

Classification collective

La classification collective exploite cette hypothèse comme étant l'information résidant dans les liens entre les données. En général, elles contiennent deux parties :

- une première sur le contenu des données qui considère les données indépendantes entre elles
- une régularisation qui permet de tenir compte des liens entre les données.

Classification itérative

Ces méthodes reposent sur l'itération de l'application de l'hypothèse de consistance. On commence, en général, par attribuer une classe à chaque donnée en fonction de ses caractéristiques. Ensuite, on classifie en fonction des voisins en respectant les deux parties de l'hypothèse présentée ci-dessus. On répète alors cette étape jusqu'à stabilisation des classes.

Cette description sommaire résume le principe de l'algorithme de classification itérative. Le problème de ce dernier est l'impossibilité de prévoir la convergence vers l'équilibre des classes. Une amélioration possible est d'utiliser l'échantillonnage de Gibbs afin d'estimer l'équilibre après un certain nombre d'itérations.

Limites

Ces algorithmes considèrent des données statiques dans le temps. Si la structure ou le contenu des données change, il faut recommencer les itérations depuis le début. Ceci rend ces méthodes difficilement adaptables aux données dynamiques, même s'il reste possible d'utiliser des graphes résumés[1] ou d'envisager de créer une relation artificielle entre les données correspondant au temps. Il est donc intéressant de chercher à développer un cadre qui s'adapte à de telles données et exploite les relations entre données.

Enfin, le fait de considérer le graphe entier pour réaliser la classification demande de grandes ressources calculatoires et considérer des paquets[2] permet d'esquiver ce problème. De plus, il est difficile d'exprimer l'influence mutuelle de données éloignées dans le graphe.

[1] Voir l'étude bibliographique réalisée en section 3.2.3

[2] Le modèle que j'ai développé, comme exposé en section 3.3.4

C’est une motivation supplémentaire pour la classification par paquets qui permet d’isoler de petits groupes de données dans lesquels on exprime pleinement les relations.

3. DEROULEMENT DU STAGE

3.1 PROBLEMATIQUE

Le domaine de l'apprentissage automatique a suivi une évolution allant de données indépendantes et identiquement distribuées à des données liés d'abord par un type de relations, puis par plusieurs.

3.1.1 Données relationnelles et dynamique

Parmi les multiples relations qui peuvent exister une semble poser problème : le temps. Si le temps semble porteur d'information que l'on pourrait utiliser pour améliorer ses résultats, la façon dont il l'apporte n'est pas claire : est-ce dans la durée ? Dans la succession d'événements ? Dans la distance temporelle ?

Aujourd'hui une littérature se développe autour de cette temporalité. Elle rassemble plusieurs communautés, des mathématiciens qui travaillent sur les graphes et voit dans les réseaux sociaux une application des modèles qu'ils avaient développés (Cortes, 2003), des statisticiens qui adaptent des méthodes de contrôle d'erreur (McCulloh, 2008), des adeptes du clustering (Chi, Song, Zhou, Hino, & Tseng, 2009) qui essaient de faire evoluer leur partionnement au cours du temps.

La communauté de l'apprentissage automatique commence par rajouter le temps comme une information dans des tâches de classification (Sharan, 2008). Mais pour l'instant les tâches proposées sont très spécifiques, et généralement très dépendantes des données auxquelles ont accès les laboratoires et il est alors difficile de comparer les résultats présentés.

Dans ce contexte, la problématique de mon stage est l'exploration de moyen d'exprimer les relations entre données et notamment un début d'exploitation de la dynamique.

3.1.2 Objectifs

Derrière cette tâche d'exploration se présente trois objectifs.

Bibliographie

La réalisation d'une étude bibliographique qui permet de cerner les problématiques du domaine et qui doit être prolongée au cours du stage afin de se tenir informé des derniers articles concernant le sujet, notamment avec la tenue de conférence comme ICML ou KDD.

Formalisation

La formalisation d'un modèle général qui permet d'être appliqué à plusieurs problèmes, sous-instance de la problématique générale de l'exploitation du temps.

Expérimentation

La réalisation d'expériences permettant de présenter le modèle. Ces expériences doivent présenter des résultats clairs afin de convaincre d'autres chercheurs de la validité et de l'intérêt du modèle.

3.1.3 Enjeux

Comme nous l'avons présenté plus haut, la question de la temporalité est nouvelle dans le domaine de la classification de données relationnelles et pour le laboratoire l'intérêt de mon travail est donc tout d'abord de pouvoir se tenir informé du travail d'autres centres de recherches et d'en observer les atouts et les limites.

Cette tâche est gourmande en temps et voir un stagiaire la réaliser est intéressant car cela permet de ne pas rajouter cette charge aux chercheurs ou doctorants qui peuvent alors continuer leur travail sur des projets en cours et, si jamais la piste de l'exploitation de la temporalité ne se révèle pas si intéressante à postériori, le laboratoire peut décider de ne pas commencer de projets à long termes pour l'exploiter.

Un autre enjeu est apparu au cours du déroulement du stage. Pour l'instant aucun laboratoire ou chercheur ne semble se démarquer sur la question de la temporalité, alors que des grands noms existent dans le domaine : Mitchell, Getoor, Zhu pour les noms et Canergie Mellon University, University of Maryland pour les centres de recherche. Si le laboratoire arrive à proposer rapidement un modèle général, même simple, il peut espérer y gagner en reconnaissance et en notoriété.

D'un point de vue personnel, travailler sur une tâche nouvelle me permet d'en un premier temps d'approfondir mes connaissances en apprentissage automatique car la lecture des derniers articles parus nécessite une bonne compréhension du domaine et j'ai donc du lire des articles plus anciens ou des livres de cours afin de compenser mes carences initiales. Cette tâche étant intéressante pour les entreprises, comme nous le verrons dans la section 4.2.1, et le laboratoire ayant de nombreux liens avec des partenaires industriels, m'y intéresser semblait en accord avec mon projet professionnel et me permis de prendre contact avec Thales pour une thèse CIFRE.

3.2 ETUDE BIBLIOGRAPHIQUE

La bibliographie permet d'étudier un domaine au travers des multiples publications scientifiques disponibles le concernant. Afin de ne pas se noyer dans cette masse, il faut garder en tête l'objectif de l'étude et suivre une certaine méthodologie.

3.2.1 Objectifs

Comme nous l'avons présenté dans la partie 0, la bibliographie que j'ai réalisée devait donner une vue d'ensemble des problèmes concernant la dynamique des données autour du domaine de l'apprentissage artificiel. Je devais donc réaliser un inventaire d'articles en prenant en notant des points importants comme le problème traité et se demander s'il s'agit d'une instance d'une problématique large ou d'un cas particulier, ou les données utilisées pour savoir si elles sont disponibles et possèdent un caractère temporel

Une fois un éventail significatif d'articles collecté, je devais présenter les résultats à mon encadrant, Patrick Gallinari, pour savoir s'il y avait une problématique ou une tâche qui en émergeait et un état de l'art à améliorer. Il est apparu que, contrairement à des problématiques comme le traitement de la langue naturelle qui est constituée de tâches classiques comme la reconnaissance d'entités nommées, la question de la dynamique n'en avait pas encore.

Le troisième objectif est de suivre constamment les publications sur le sujet afin de suivre les idées des autres centres de recherche.

3.2.2 Méthodologie

La majeure difficulté d'une étude bibliographique est de ne pas se noyer dans la masse de publications disponibles grâce à des sites internet comme Google Scholar ou ACM Portal[3].

[3] Respectivement http://scholar.google.fr et http://portal.acm.org/portal.cfm. D'autres existent (arXiv, Citeseer,...) mais j'ai principalement utilisé ces deux là.

Mots clés

La première méthode de recherche est d'utiliser des requêtes composées de mots clés comme *temporal*, *classification*, *learning*, … qui fourniront une vague de résultats que l'on peut alors trier soit par date soit par nombre de citations suivant que l'on recherche de la nouveauté ou du reconnu. Mais procéder comme tel présente deux problèmes. Premièrement, ce n'est pas efficace car la précision des résultats est à priori très faible. Enfin, en utilisant mal les mots clés il est possible de s'éloigner fortement du domaine d'origine, surtout lorsque l'on cherche à traiter de la dynamique, le temps est présent dans de nombreux domaines, et de méthodes d'apprentissage automatique qui sont utilisés là aussi dans d'autres domaines.

Conférences

Une première solution consiste à définir des conférences qui s'intéressent au sujet. Elles sont un lieu d'échange privilégié entre les différents centres de recherches (cf. partie 4.2.2) et sont organisées autour de thèmes. Patrick Gallinari m'a conseillé la liste suivante comme une base de travail :

- KDD : Knowledge Discovery and Data-mining;
- WWW : International World Wide Web Conference;
- ICML : International Conference on Machine Learning;
- CIKM : Conference on Information and Knowledge Management;
- ICWSN : International Conference on Weblogs and Social Media;
- ASONAM : international conference on Advances in Social Network Analysis and Mining.

Ces conférences proposent en général sur leur site internet une liste des articles acceptés et qui seront présentés.

Citations

Lors de la rédaction d'un article scientifique, il est important d'en présenter le contexte et les connections avec les résultats d'autres équipes scientifiques, ce qui ce fait en citant leurs articles. Il existe donc un réseau de relation entre les articles qui peuvent être cités pour plusieurs raisons :

- les citations explicatives : l'équipe qui rédige s'appuie sur le résultat d'un autre travail et cite donc ce travail, cette référence est donc importante à lire si on veut comprendre le fond de l'article ;
- les citations contextuelles : afin de présenter le contexte de son travail, l'équipe qui rédige l'article présente les résultats d'autres équipes sur des thématiques proches. L'intérêt de cette dépendance peut varier en fonction de l'intention du citeur :
    - s'il y a effectivement convergence entre les deux articles, il peut être intéressant de consulter la référence pour trouver d'autres problématiques ;
    - sinon l'équipe peut ne citer un article que pour élargir le champ d'application de son article à des domaines, souvent porteurs de crédits, comme la bioinformatique, les réseaux sociaux ou l'informatique verte ;

En suivant les directions indiquées par les citations, il est possible de naviguer au sein de portails comme ceux présentés et de ne pas trop se perdre.

Méthode

Une bonne méthodologie pour trouver des articles parlant d'une problématique est donc de dresser une liste de conférences qui traitent du sujet, qui permet d'établir un premier ensemble d'articles. Cet ensemble est ensuite étendu grâce au jeu des citations. Une fois ce travail accompli, et les articles lus, il est possible de dégager un certain nombre de mots clés pertinents qui permette d'élargir une dernière fois sa base d'articles.

3.2.3 Constat

Si l'on en croit les publications de ces trois dernières années, la communauté de l'apprentissage artificiel commence à explorer les différentes possibilités qu'offre la prise en compte du temps. Je présente ici deux principales utilisations du temps : comme une distance qui pondère d'autres relations et comme une contrainte pour assurer la continuité d'un modèle.

Le temps comme distance

On reprend ici la représentation des données relationnelles sous forme de graphe, comme présenté dans la partie 2.2.3. Le poids associé à une arête influence généralement la classification des nœuds de la façon suivante : plus il est élevé, plus les nœuds ont de grandes chances d'être de même classe. L'idée de cette utilisation du temps est de considérer qu'il a une influence sur le poids des arêtes : un lien récent aura un poids plus fort qu'un lien ancien.

Afin de mettre en place cette idée, il faut introduire un formalisme (Cortes, 2003) qui permet de créer et mettre à jour à chaque pas de temps, un graphe dont les arêtes sont pondérées par leur ancienneté.

Graphe résumé

On assigne un poids à toute arête dans le graphe et, par convention, une arête de poids nul équivaut à l'absence d'arête. On définit la somme de deux graphes G_1,G_2comme étant le graphe G dont :

- l'ensemble des nœuds est l'union des ensembles des nœuds des deux graphes G_1,G_2 ;
- le poids d'une arête de G est la somme des poids de cette arête dans G_1et G_2.

Afin de suivre l'évolution des données dans le temps, à chaque pas de temps t on réalise un graphe g_t qui en contient l'activité : nœuds ou arêtes créés ou détruits.

On introduit alors la notion de graphe résumé à l'instant T comme étant la somme pondérée des graphes des pas de temps précédent T :

$$G_T = \sum_{t=1}^{T} \omega_t g_t \ , avec \sum_{t=1}^{T} \omega_t = 1$$

Reste à choisir les ω_i pour exprimer l'effacement progressif des liens les plus anciens. Pour cela, on choisit :

$$\forall i \in \{1, \dots, T\}, \qquad \omega_i = \theta^{t-i}(1-\theta), \qquad avec\ 0 \leq \theta \leq 1$$

Ce qui permet de reformuler l'expression de G_T :

$$G_T = \theta G_{T-1} + (1-\theta) g_t$$

Si l'article de Cortes (Cortes, 2003) présente le modèle mathématique des graphes résumés, il est possible de l'utiliser (Sharan, 2008) pour des tâches de classification de documents.

Distance temporelle

Le problème d'un tel graphe résumé et qu'il peut contenir un lien qui ne respecte aucune causalité. On peut alors définir une notion de distance temporelle (Tang, 2009) qui compte le nombre de pas de temps nécessaire pour établir un lien entre deux nœuds du graphe.

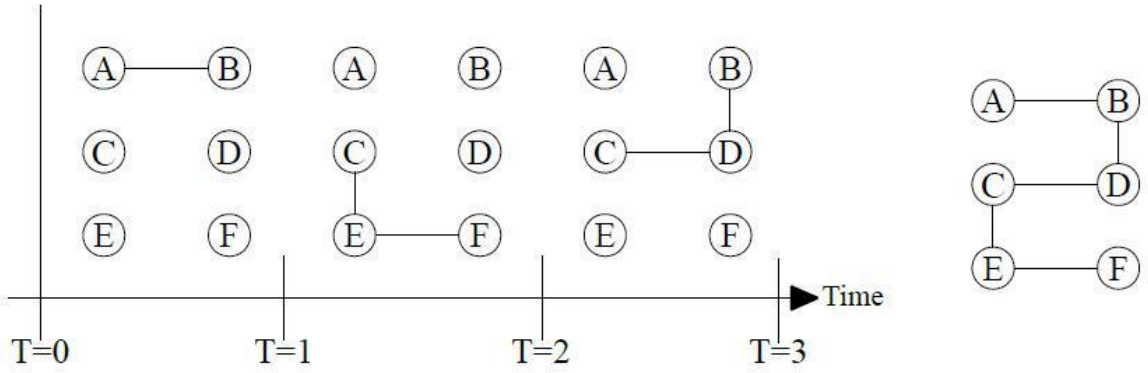

Figure 4 : exemple de graphe temporel à gauche et son résumé à droite (Tang, 2009)

Comme on peut le voir Figure 4, le résumé temporel laisse sous entendre un lien entre A et F, qui n'existe en réalité pas. Le modèle de distance temporel aura assignée une distance infinie entre A et F pour permettre de respecter la causalité.

Si cette notion semble intéressante, je ne l'ai pour l'instant pas vue appliquée dans un modèle de classification.

Le temps comme contrainte

Cette vision est généralement retrouvé dans le domaine du *clustering*, qui est une forme d'apprentissage non-supervisé. Pour ce domaine, l'intérêt est de pouvoir suivre l'évolution temporelle des communautés créées par les algorithmes utilisées. Mais ces derniers sont très sensibles aux changements dans le graphe : une arête qui disparait peut modifier radicalement les communautés.

Le temps est alors apparu comme une contrainte qui est rajoutée afin que les modifications d'un pas de temps à l'autre reste minimes (Chi, Song, Zhou, Hino, & Tseng, 2009). On utilise alors les partitions trouvées à un pas de temps comme base pour créer les partitions de l'instant suivant. Une manière claire d'écrire ce problème introduit la notion de coût de capture CS et coût temporel CT (Lin, 2008). Le premier mesure la compatibilité entre une partition trouvée et la structure du graphe, au pas considéré. Le second mesure la variation entre la partition du pas précédent et celle que l'on vient de trouver. On définit alors un coût total :

$$cost = \alpha . CS + (1-\alpha) . CT$$

Cette approche est aussi utilisée pour des modèles à variable latentes, pour le même objectif : limiter l'impact des changements de structures afin que les déplacements des objets dans l'espace latent ne soient pas trop importants (Fu, 2009).

Manque d'article de référence

Comme précisé en 3.1.1, si beaucoup de communautés s'intéressent à la question, il n'est pour l'instant pas possible de dégager une tâche type sur laquelle un état de l'art pour s'établir et des modèles pourraient être comparés.

Une des raisons de ce problème est la rareté des données. On reparlera de ce problème durant la dernière partie de ce rapport, qui est issu de conflits de propriétés : les entreprises, notamment Facebook et Twitter, possèdent de gigantesques bases de données mais ne peuvent pas les publier car elles représentent une partie de leur fond de commerce. Toujours est-il que les données libres qui possèdent une information temporelle sont rares. Certaines qui auraient pu convenir, comme celles de la coupe KDD 1999[4] qui propose un challenge de détection d'attaques sur un réseau, ont été traitées pour correspondre à l'usage de l'époque et l'information temporelle, alors non utilisée, a disparue.

Une autre est la difficulté à exprimer mathématiquement l'apport du temps dans un problème de classification. Deux façons valables ont été présentées, et il est possible d'en trouver d'autres, mais toutes deux dépendent d'un besoin et s'appliquent à un problème particulier. Trouver une formulation générale de l'apport du temps semble hors de portée.

Enfin, la dynamique des données pose plusieurs problèmes (Masud, 2010) : les concepts peuvent changer dans le temps : de nouveaux apparaissent ou d'anciens dérivent ; les caractéristiques significatives des objets peuvent évoluer et, si l'on considère des flux de données, ces derniers sont potentiellement infinis.

La problématique présentée en 3.1.1, conclut cette étude bibliographique : il est nécessaire d'explorer différents moyens d'exploitation des relations entre données et d'intégrer petit à petit le temps.

3.3 FORMALISATION D'UN MODELE

La formulation d'un modèle est la suite de ce travail exploratoire, ce modèle ayant pour but de répondre à la problématique dégagée par l'étude bibliographique. Après avoir discuté avec Patrick Gallinari et Ludovic Denoyer, l'idée de la classification par paquets est apparue comme une bonne piste.

3.3.1 Enjeu et difficulté

Le but de cette étape est de concrétiser le travail exploratoire. Un modèle simple et fonctionnel avec des résultats corrects justifierait l'intérêt en tant que problème de recherche de la problématique dégagée après l'étude bibliographique. D'un point de vue personnel, il s'agit de montrer et renforcer ma compréhension des concepts de l'apprentissage automatique.

La formalisation d'un problème, alors que la formation que j'ai suivie m'a surtout appris à en résoudre, est une tâche difficile. Pour y arriver, il faut selon moi trois qualités qui sont étroitement liés :

- une rigueur dans l'expression des concepts afin de ne pas les mélanger ;
- une connaissance générale du domaine et un recul sur ses problématiques ;
- l'habitude de cette tâche.

Si mon cursus m'a appris la première, je n'avais pas suffisamment de recul et d'habitude pour arriver à formaliser seul, depuis le début, un modèle. C'est donc en discutant à de nombreuses reprises avec des chercheurs, dont Patrick Gallinari, mon encadrant, et Ludovic Denoyer, de l'équipe MALIRE, que nous avons abouti à l'idée et au début de la formalisation

[4] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
http://kdd.ics.uci.edu/databases/kddcup99/task.html

de l'apprentissage par paquets, dans le cas d'un paquet de taille deux. J'ai ensuite écrit le problème dans le cas général.

3.3.2 Motivation

L'exploration de la classification par paquets est motivé par deux idées entrevues lors de la présentation du domaine de l'apprentissage statistique, partie 2.2.3.

La première est de proposer un modèle souple qui permettra d'exploiter la dynamique des données en même temps que les relations entre ces dernières. Le modèle par paquet permet de faire intervenir le temps de deux façons. A la création des paquets, que l'on peut envisager par exemple comme une séquence d'événements proches dans le temps. Cela s'inspire du passage de la classification de données vectorielles à celle de données séquentielles et revient à supposer que la chronologie des événements est porteuse d'information. Lors de la phase de régularisation, en faisant intervenir un poids temporel.

Cette souplesse peut permettre d'envisager un début d'exploitation du temps dans la classification car elle permet d'en contenir plusieurs aspects présentés durant la conclusion de la bibliographie, partie 3.2.3, à savoir la durée et la succession chronologique.

La seconde est de permettre de se concentrer sur un sous-ensemble du très grand, en général, ensemble des données d'un problème. L'idée derrière la classification par paquet vient du problème de la détection d'anomalies : peut-être que considérer des données similaires par paquet permettrait de meilleurs résultats. La détection d'intrusion, où une succession de traceroute puis map[5] vers une même machine ou issus d'une même machine peuvent être suspects alors que les commandes en elles-mêmes sont normales, consolide cette intuition. Les résultats de mes expériences sur la base de données CORA (cf. partie 3.4.3) aussi.

En se concentrant sur un ensemble plus petit de données on peut exprimer des relations particulières difficiles à exprimer sur le graphe en entier. De plus cela limite la demande en ressources de calculs, ce qui n'est pas négligeable dans le cas de grandes bases de données.

3.3.3 Notations

On note $\mathcal{X}$ l'ensemble des données et $X_P = (x_i)_{1 \leq i \leq P}$ un paquet de taille P, c'est-à-dire un sous-ensemble de $\mathcal{X}$ où x_i représente le vecteur de contenu de l'élément.

L'ensemble des étiquettes possibles est noté $\mathcal{Y}$. Le but de la classification collective par paquets est de trouver un étiquetage $Y_P = (\hat{y}_i)_{1 \leq i \leq P}$pour chaque paquet X_P. On traite ici des problèmes de classification binaire et on a donc $\mathcal{Y} = \{-1,1\}$.

On note $f : \mathcal{X} \to [-1,1]$ un classificateur sur le contenu de chaque élément.

On note $\mathcal{H}_\gamma = \mathcal{Y}^P \to [0,+\infty[$ une fonction de régularisation dont le but est de prendre en compte les relations entre les données.

Enfin, on note Δ une fonction de coût comme la distance euclidienne.

3.3.4 Classification par paquets

On formule maintenant le problème d'apprentissage suivi de celui de l'inférence dans le cadre de la classification par paquets. Les deux peuvent être vus comme un problème d'optimisation. Le but est d'être capable d'assigner un score à chaque donnée, compris entre moins un et un ici, pour ensuite décider de sa classe avec plus ou moins de confiance.

[5] Commandes qui permettent de connaître le chemin vers une machine dont on connaît l'adresse IP (traceroute) et de connaître les ports ouverts (map).

Apprentissage

Le cadre de l'apprentissage semi-supervisé permettra, car il a été en partie développé pour cela (Chapelle, 2006), de dépasser la rareté des données étiquetées, mais par souci de simplicité, on choisit le cadre supervisé pour formulé le problème. On dispose donc d'un ensemble d'entrainement noté :

$$X_T = \left(X_{P_j}\right)_{1 \le j \le T} = \left(\left(x_{i_j}\right)_{i \le i \le P_j}\right)_{1 \le j \le T}$$

On connait pour chaque donnée x_{i_j}sa classe réelle, notée y_{i_j}.

Le problème s'écrit comme la minimisation de l'erreur entre le classificateur sur le contenu et l'étiquette connue (ce qui revient à considérer les données comme indépendantes et identiquement distribuées) et de la contrainte qui permet de prendre en compte les dépendances.

$$\min\left(\sum_{j=1}^{T}\left(\sum_{i_j=1}^{P} \Delta\left(y_{i_j}, f\left(x_{i_j}\right)\right) + H_\gamma\left(f\left(x_{1_j}, x_{2_j}, \dots, x_{P_j}\right)\right)\right)\right)$$

Ce qui revient, puisque tous les termes sont positifs, à minimiser sur chaque paquet :

$$\forall \mathrm{j} \in [\![1, \mathrm{T}]\!], \qquad \min\left(\sum_{i_j=1}^{P} \Delta\left(y_{i_j}, f\left(x_{i_j}\right)\right) + H_\gamma\left(f\left(x_{1_j}, x_{2_j}, \dots, x_{P_j}\right)\right)\right)$$

Inférence

Une fois f appris, on peut l'utiliser pour l'inférence. Il s'agit de minimiser cette fois de trouver la famille $\widehat{y_P} = \left(\hat{y}_{i_j}\right)_{1 \le i \le P_j}$ qui minimise l'écart avec le score donné par le classificateur f pour chaque donnée et la contrainte de régularisation.

$$\widehat{y_P} = argmin_{\left(\hat{y}_{i_j}\right)_{1 \le i_j \le P_j}}\left(\sum_{i_j=1}^{P_j} \Delta\left(\hat{y}_{i_j}, f\left(x_{i_j}\right)\right) + H_\gamma\left(\hat{y}_{1_j}, \dots, \hat{y}_{P_j}\right)\right)$$

3.3.5 Régularisations

La régularisation H_γ peut prendre différentes formes afin de s'adapter aux relations entre les données, on en présente ici deux : une naïve et une utilisant un poids structurel.

Régularisation naïve

Cette régularisation considère que tous les éléments d'un paquet sont connectés par une même relation, celle d'appartenir au paquet en question, et considère que des éléments d'un même paquet sont susceptibles d'avoir la même étiquette.

Même si cette régularisation apparait naïve, on peut penser à un champ d'application : la classification de documents. On considère un nombre de sources d'informations sur des sujets communs, comme par exemple les flux dédiés aux actualités françaises de journaux comme Le Monde et Libération. Un paquet est commencé du dernier article publié par la source 1, la source 2, … la source P.

Dans ce cas l'expression de H_γ est :

$$H_\gamma\left(y_{1_j}, \dots, y_{P_j}\right) = \gamma . \sum_{i_j=1}^{P_j} \sum_{k_j=\, i_j+1}^{P_j} \Delta(y_{i_j}, y_{k_j})$$

On considère maintenant que le coût est la distance euclidienne. Le problème d'inférence s'écrit alors :

$$argmin_{\left(\hat{y}_{i_j}\right)_{1\le i_j\le P_j}} \left(\sum_{i_j=1}^{P_j} \left(\hat{y}_{i_j} - f\left(x_{i_j}\right)\right)^2 + \gamma . \sum_{i_j=1}^{P_j} \sum_{k_j=\, i_j+1}^{P_j} \left(\hat{\mathrm{y}}_{\mathrm{i_j}} - \hat{\mathrm{y}}_{\mathrm{k_j}}\right)^2 \right)$$

On note :

$$C_j\left(\left(\hat{y}_{i_j}\right)_{1\le i_j\le P_j}\right) = \sum_{i_j=1}^{P_j} \left(\hat{y}_{i_j} - f\left(x_{i_j}\right)\right)^2 + \gamma . \sum_{i_j=1}^{P_j} \sum_{k_j=\, i_j+1}^{P_j} \left(\hat{\mathrm{y}}_{\mathrm{i_j}} - \hat{\mathrm{y}}_{\mathrm{k_j}}\right)^2$$

La fonction C est infiniment dérivable et continue. Son gradient vaut :

$$\forall i \in \{1, \dots, P_j\}, \qquad \left(\frac{\partial C}{\partial y_{k_j}}\right)\left(\left(\hat{y}_{i_j}\right)_{1\le i_j\le P_j}\right) = 2\left(\hat{y}_{k_j} - f\left(x_{k_j}\right)\right) + 2\gamma . \sum_{\substack{m_j \ne k_j \\ m_j = 1_j}}^{P_j} \left(\hat{y}_{k_j} - \hat{y}_{m_j}\right)$$

La matrice hessienne est donc de la forme :

$$H(C_j) = 2. \begin{pmatrix} 1-\gamma P_j & -\gamma & \cdots & -\gamma \\ -\gamma & \ddots & & \vdots \\ \vdots & & & -\gamma \\ -\gamma & \cdots & -\gamma & 1+\gamma P_j \end{pmatrix}$$

Cette matrice est définie positive pour $\gamma \in]-\frac{1}{P_j}, +\infty[$.

Comme γ permet de réaliser l'équilibre entre importance du contenu ($\gamma \to 0$) et importance de la régularisation ($\gamma \to \infty$) on considère en général $\gamma \in [0,1]$. La fonction C est donc convexe.

Le problème de minimisation exposé plus haut revient à la résolution du système algébrique :

$$H(C_j).\hat{Y}_j = F_j, \qquad o\grave{u}\ \hat{Y}_j = \begin{pmatrix} \hat{y}_{1_j} \\ \vdots \\ \hat{y}_{P_j} \end{pmatrix}, F_j = \begin{pmatrix} f_{1_j} \\ \vdots \\ f_{P_j} \end{pmatrix}$$

Comme la taille du paquet est limitée, on peut envisager d'inverser la matrice hessienne pour résoudre ce système algébrique.

Régularisation sur la structure

Cette régularisation considère reprend la méthode naïve en ajoutant une notion de poids structurel qui pondère le lien entre deux données d'un même paquet. L'idée est de considère que des messages échanges par des communautés proches sont susceptibles d'être étiquetés de manière semblable. Cette hypothèse peut-être intéressante dans le cadre de la tâche de classification de document suivante : on considère une boîte de réception de courriel avec des dossiers créés par l'utilisateur et on essaie d'apprendre à classer les courriels dans les bons dossiers.

Pour faciliter la compréhension, on considère ici le cas où les données sont des messages entre personnes. La structure de la donnée i_j est alors représentée par un vecteur :

$$s_{i_j} = \left(s_{i_j}^1, \dots, s_{i_j}^m\right)\ o\grave{u}\ s_{i_j}^k = \begin{cases} 1\ si\ la\ personne\ k\ prend\ part\ \grave{a}\ la\ communication \\ 0\ sinon \end{cases}$$

On définit alors un poids structurel (et une notation abusive) :

$$p_s\left(s_{i_j}, s_{k_j}\right) = p_{i_j,k_j} = \frac{< s_{i_j}, s_{k_j} >}{\left\| s_{i_j} \right\| \left\| s_{k_j} \right\|}$$

On remarque que, comme le produit de réels et le produit scalaire sont symétriques, la fonction de poids est symétrique : $p_{i_j,k_j} = p_{k_j,i_j}$. De plus, le résultat est un réel compris entre 0 et 1.

Et on modifie l'expression de la régularisation pour en tenir compte :

$$H_\gamma\left(y_{1_j}, \dots, y_{P_j}\right) = \gamma . \sum_{i_j=1}^{P_j} \sum_{k_j=i_j+1}^{P_j} p_{i_j,k_j} . \Delta(y_{i_j}, y_{k_j})$$

On réécrit le problème d'optimisation correspondant à l'inférence :

$$argmin_{\left(\hat{y}_{i_j}\right)_{1\leq i_j\leq P_j}}\left(\sum_{i_j=1}^{P_j}\left(\hat{y}_{i_j}-f\left(x_{i_j}\right)\right)^2+\gamma.\sum_{i_j=1}^{P_j}\sum_{k_j=\,i_j+1}^{P_j}p_{i_j,k_j}.\left(\hat{\mathrm{y}}_{\mathrm{i_j}}-\hat{\mathrm{y}}_{\mathrm{k_j}}\right)^2\right)$$

On note :

$$C_j\left(\left(\hat{y}_{i_j}\right)_{1\leq i_j\leq P_j}\right)=\sum_{i_j=1}^{P_j}\left(\hat{y}_{i_j}-f\left(x_{i_j}\right)\right)^2+\gamma.\sum_{i_j=1}^{P_j}\sum_{k_j=\,i_j+1}^{P_j}p_{i_j,k_j}.\left(\hat{\mathrm{y}}_{\mathrm{i_j}}-\hat{\mathrm{y}}_{\mathrm{k_j}}\right)^2$$

De même que pour la régularisation naïve, la fonction est indéfiniment dérivable et continue. Son gradient vaut :

$$\forall i\in\{1,\dots,P_j\},\qquad\left(\frac{\partial C}{\partial y_{k_j}}\right)\left(\left(\hat{y}_{i_j}\right)_{1\leq i_j\leq P_j}\right)=2\left(\hat{y}_{k_j}-f\left(x_{k_j}\right)\right)+2\gamma.\sum_{\substack{m_j\neq k_j\\ m_j=1_j}}^{P_j}p_{m_j,k_j}.\left(\hat{y}_{k_j}-\hat{y}_{m_j}\right)$$

Cette expression est correcte car, comme il a été noté plus haut, la fonction de poids structurel est symétrique.

On écrit alors la matrice hessienne :

$$H(C_j)=2\begin{pmatrix}1-\gamma.\sum_{\substack{m_j\neq 1_j\\ m_j=1_j}}^{P_j}p_{m_j,1_j} & -\gamma p_{2_j,1_j} & \cdots & -\gamma p_{P_j,1_j}\\ -\gamma p_{1_j,2_j} & \ddots & & \vdots\\ & & & -\gamma p_{P_j,P_j-1}\\ \vdots & & & \\ -\gamma p_{1_j,P_j} & \cdots & -\gamma p_{P_1-1,P_j} & 1+\gamma.\sum_{\substack{m_j\neq P_j\\ m_j=1_j}}^{P_j}p_{m_j,1_j}\end{pmatrix}$$

En suivant le même raisonnement que pour la régularisation naïve et comme la fonction de poids structurel est symétrique, la matrice hessienne est au moins définie positive pour $\gamma\in[0,1]$.

La fonction C est donc convexe. Le problème de minimisation exposé plus haut revient de même que dans le cas de la régularisation (et la réflexion quant à la taille du paquet reste valable aussi) naïve à la résolution du système algébrique :

$$H(C_j).\hat{Y}_j = F_j, \qquad o\grave{u}\ \hat{Y}_j = \begin{pmatrix} \hat{y}_{1_j} \\ \vdots \\ \hat{y}_{P_j} \end{pmatrix}, F_j = \begin{pmatrix} f_{1_j} \\ \vdots \\ f_{P_j} \end{pmatrix}$$

D'autres régularisations

Il est bien entendu possible de formuler d'autres régularisations[6], notamment pour faire intervenir le temps, mais aussi pour étendre au cas de données multi-relationnelles.

Poids temporel

On peut définir une notion poids temporel qui permettra de prendre le temps en compte.

On peut par exemple penser à la formulation suivante :

$$p_t(i_j, k_j) = \frac{1}{e^{-(t_i - t_j)}}$$

Elle permet de prendre en compte la distance temporelle entre deux message i_j et k_j et permet de lier plus fortement les données proches dans le temps. Une amélioration simple est de considérer :

$$p_t(i_j, k_j) = \frac{1}{e^{-(t_i - t_j)^2}}$$

Cette expression permet de garder une symétrie dans l'expression des problèmes de minimisation d'apprentissage et d'inférence.

Une autre possibilité serait peut-être d'utiliser un classificateur à « mémoire » qui conserverait le temps associé à la dernière donnée classée dans chaque classe du problème et qui utiliserait la distance entre ces bornes temporelles et le temps de la donnée à classer.

Multi-relationnel

On peut modifier le modèle pour prendre en compte plusieurs relations dans les données en s'inspirant par exemple des travaux réalisés par Yann Jaccob, doctorant de l'équipe DAPA. L'idée principale est d'être capable de définir un poids permettant de réunir toutes les relations.

[6] Les deux présentées plus haut sont celles que j'avais développées au moment du rendu du rapport.

3.4 EXPERIMENTATION

La phase d'expérimentations vient en conclusion de la formulation du problème. Il s'agit de confronter celui-ci à une certaine réalité (les données peuvent être prétraitées). Au moment de rédiger ce rapport, j'avais écrit une plateforme, décrite en annexe 6.3, capable de réaliser la classification par paquet mais n'avait réalisé des expériences que sur la base CORA (LINQS.UMD).

Cette plateforme est développée en C++ et dépend des librairies Boost (Boost, 2010) et SQLite3 (SQLite, 2010). La première propose des améliorations comme les pointeurs intelligents et la librairie uBLAS (Boost, 2010) très intéressante pour les opérations algébrique. La seconde permet de récupérer des données que j'ai stockées dans des bases SQLite lors de leur prétraitement, ce qui était le cas pour Enron (Cohen, 2009).

3.4.1 Objectif

La première des choses est de réaliser un code correct et de vérifier son fonctionnement et de sa cohérence au modèle. Il peut être intéressant de réaliser une expérience peu significative mais dont le résultat est attendu pour cela, et c'est qui a motivé mon traitement de la base CORA (LINQS.UMD).

Il est ensuite intéressant d'appliquer son modèle à différentes instances de la problématique initiale. C'est pour cela que j'ai prévu de l'appliquer à la tâche de classification de document sur Enron (Cohen, 2009). D'autres tâches sont envisagées d'ici la fin du stage si j'arrive à les formuler correctement.

D'une manière générale, les expériences permettent de comparer ses résultats avec d'autres méthodes mais aussi de mettre en lumière certains comportements de sa méthode, ses atouts et ses faiblesses.

3.4.2 Données

Le problème principal lors de la réalisation d'expériences dans le domaine de l'apprentissage artificiel est de trouver des données. Cela est d'autant plus difficile s'il faut trouver des données étiquetées. Le risque corrélé est de formuler son modèle en fonction des données que l'on possède, démarche qui permet des résultats mais peut rendre difficile la généralisation du modèle. Il est donc en général plus intéressant de développer un modèle indépendamment de données, même si cela entraine plus de difficulté après avec peut-être la nécessité d'étiqueter « à la main » une base.

CORA

La présentation sur le site des étudiants de Lise Getoor (LINQS.UMD) est claire : il s'agit de 2708 publications scientifiques dans le domaine de l'apprentissage artificiel réparties en sept classes. Les liens entre les données sont les citations des articles, ce qui constitue 5429 relations. Le vecteur caractéristique de chaque document est le résultat de la conservation des mots significatifs du résumé de chaque texte, ici 1433 mots.

Les sept classes sont, en anglais : *Case Based*, *Genetic Algorithms*, *Neural Networks*, *Probabilistic Methods*, *Reinforcement Learning*, *Rule Learning* et *Theory*.

Représentation

Les données sont représentées en machine par deux fichiers. Le premier définit un identifiant par texte, donne sa classe et le résumé de son contenu. Chaque ligne représente un document, commence par l'identifiant puis le vecteur de caractéristiques et termine par la classe, les valeurs sont séparées par un espace. Le second définit les liaisons, chaque ligne est une citation et se présente sous la forme de l'identifiant du texte cité séparé par un espace de l'identifiant du texte citant.

Cette représentation est donc simple et il est possible de la traiter directement en C++, notamment grâce à l'objet tokenizer de la librairie Boost, qui permet de séparer des chaînes de caractères.

Prétraitement

Le but de mon expérience avec CORA était de montrer que mon code était correct et correspondait au modèle avec régularisation naïve. Pour cela, il était nécessaire de trouver des données étiquetées présentant la particularité que demande cette régularisation, à savoir que les données d'un même paquet ont la même étiquette.

J'ai donc prétraité la base pour la rendre compatible avec cette hypothèse en utilisant le langage Python. J'ai tout d'abord choisi une classe parmi les sept, en l'occurrence *Theory*, pour l'opposer à toutes les autres et me rapporter au problème de classification binaire. J'ai ensuite ordonné le fichier contenant les structures de façon à faire apparaitre en premier les exemples de la classe choisie.

Ainsi, lors de la lecture est de la création des paquets par ma plateforme en C++, comme elle lisait les données dans l'ordre du fichier, ces dernières arrivaient par paquets de données de même classe à l'exception d'au pire un, qui faisait transition.

Enron

Enron était une entreprise texane qui connut une fin tumultueuse et dont les communications internes ont été rendues publiques. Elles sont désormais utilisables pour des tâches de classification, notamment grâce à William W. Cohen (Cohen, 2009). Certains messages ont pu être retirés conformément au souhait des personnes concernées. Il reste néanmoins plus de cinq cent mille mails (avec des doublons) au format MIME. Les destinataires, sujets, dates, contenus et autres champs sont disponibles.

Représentation

La base de données consiste en un répertoire, nommé sobrement maildir, qui contient les boîtes mails des employés de la société. Ces boîtes contiennent à leur tour des dossiers classiques comme « Envoyés », « Poubelle » ou des dossiers créés par des utilisateurs. Ces dossiers contiennent des fichiers qui représentent le mail au format MIME. Un exemple est présent en annexe, cf. 6.2

Prétraitements

Cette base est donc inutilisable directement comme beaucoup de bases de données réelles. Je l'ai donc prétraitée, là aussi grâce au langage Python qui contient de nombreuses librairies et notamment une pour gérer les emails au format MIME.

Ce prétraitement visait à :

- donner un identifiant à chaque compte mail présent dans les messages ;
- conserver les structures des messages ;
- donner un identifiant à chaque mail ;
- traiter le contenu et le sujet de chaque mail pour ne garder que les mots les plus significatifs ;
- enregistrer le tout dans une base de donnée relationnelles grâce à SQLite.

Cette base de données comprend cinq tables :

- enron (id, date, structure_id, subject_id, payload_id) :
    - id est l'identifiant du message ;
    - date est le nombre de seconde écoulées depuis le premier message du corpus ;

    - structure_id est l'identifiant du vecteur de structure du mail, à savoir l'émetteurs et les destinataires, que l'on retrouve dans la table structure ;
    - subject_id est l'identifiant du vecteur résumant les mots du sujet du mail que l'on retrouve dans la table subject
    - payload_id est l'identifiant du vecteur résumant les mots du contenu du mail que l'on retrouve dans la table payload
- structure (id, key, value) est une représentation éparse du vecteur de structure :
    - id est l'identifiant du vecteur de structure ;
    - key correspond à l'identifiant d'un compte ;
    - value vaut -1 si la personne émet, 1 sinon ;
- subject (id,key,value) est une représentation éparse du vecteur de contenu du sujet du mail:
    - id est l'identifiant du vecteur de contenu du sujet ;
    - key correspond à l'identifiant d'un mot ;
    - value vaut 1 si le mot est présent ;
- payload (id,key,value) est une représentation éparse du vecteur de contenu du corps du mail :
    - id est l'identifiant du vecteur de contenu du corps ;
    - key correspond à l'identifiant d'un mot ;
    - value vaut 1 si le mot est présent ;

On peut ensuite lire cette base de données depuis la plateforme en C++, langage pour lequel SQLite3 existe et est documentée.

Enron Flat

Ron Bekkerman propose sur sa page personnelle (Bekkerman) une version prétraité de la base d'Enron qui ne conserve que sept boîtes uniquement constituées des dossiers créés par les utilisateurs et enlève les doublons.

Afin de réaliser la même tâche, à savoir apprendre à répartir automatiquement les mails dans les dossiers utilisateurs, j'envisage de réaliser le même formatage vers SQLite de cette base.

Cette tâche me paraît intéressante car Ron Bekkerman avoue lui-même que prendre en compte les relations entre les données peut se révéler intéressant, mais qu'il n'avait pour l'instant, en 2004, vu personne le faire. Le modèle de la classification par paquets devrait être capable d'améliorer ses résultats.

3.4.3 Résultats

La seule expérience dont je peux présenter actuellement des résultats est celle réalisée sur CORA pour, je le rappelle, vérifier le fonctionnement correct de ma plateforme. Il s'agissait donc de données arrivant par paquets composés d'individus de même classe et en utilisant le modèle à régularisation simple.

La valeur de γ est ici 1.

Les diagrammes de résultats Figure 5 montrent la précision, le rappel et le F1-score comme définis en partie 2.2.1.

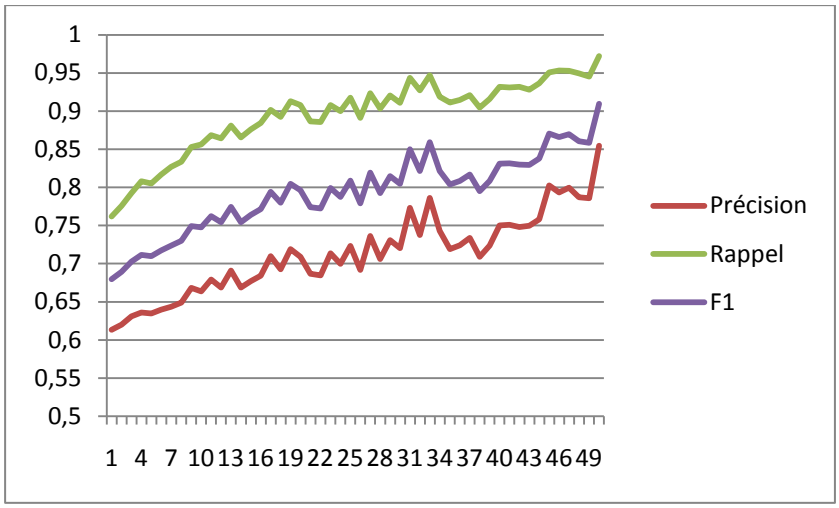

Figure 5 : résultats sur CORA, l'abscisse indique la taille des paquets

Comme on peut le voir, la méthode de classification par paquet fournit de très bons résultats en rappel et des résultats corrects en précision et donc de bons résultats en F1.

Le point le plus important est que l'utilisation de paquets augmente la précision. Cela confirme que la régularisation fonctionne.

Le rapport technique publié par Lise Getoor (Sen, 2007) présente des résultats en F1-score de 0,84 dans un contexte multi-classe. Mon modèle est à priori de performance comparable. On peut encore améliorer ses résultats en s'autorisant une valeur de γ supérieure à un. Cela signifie que l'on donne plus d'influence à la régularisation qu'au classificateur sur le contenu.

Il faut tout de même rester prudent sur les conclusions. Premièrement je réalise une tâche de classification binaire et pour comparer définitivement avec les résultats de Sen et Getoor, il faudrait réaliser une classification multi-classe, à savoir réaliser la même expérience avec les autres classes sur chaque paquet et assigné la classe dont le score aurait été le plus élevé. Ensuite, les données étaient prétraitées de façon à correspondre à une hypothèse précise.

Cette expérience m'a tout d'abord permis de vérifier la plateforme que j'ai développée. Elle permet de confirmer ensuite que la régularisation naïve permet bien d'améliorer les résultats en classification par rapport au contenu seul si les données correspondent à l'hypothèse qu'elle effectue.

Il ne faut à priori pas donner plus de portée que cela à ces résultats et attendre d'autres expériences pour conclure quand à l'apport réel en performances du modèle.

3.4.4 Prochaines expériences

La tâche que je prépare alors que je rédige ce rapport est une tâche de classification de documents sur le corpus Enron Flat.

Elle consiste à apprendre comment classer automatiquement les messages dans les dossiers créés par un utilisateur et a été réalisée par Ron Bekkerman (Bekkerman). Comme ce dernier l'indique il n'a pas utilisé la structure de graphe du réseau de communication inhérente aux mails.

Je pense que l'utilisation de la régularisation avec poids structurel fournira de bons résultats. Cette intuition s'appuie sur deux idées :

- les dossiers utilisateurs représentent deux choses en général. Des conversations autour d'un sujet ou avec une personne, un groupe de personnes en particulier ;
- dans le contexte d'une entreprise, les discussions autour d'un sujet concernent généralement un département soit une communauté de personnes.

Le modèle par paquets utilisant les communautés et le contenu pour classer devrait parvenir à de bons résultats, en classification binaire dans un premier temps.

Pour poursuivre il peut être intéressant de passer au contexte multi-classe et de reprendre les deux expériences, sur CORA et Enron Flat.

J'ai ensuite envisagé de continuer sur deux tâches :

- la détection d'anomalie dans Enron, la compagnie ayant connu des tumultes judiciaires, la base connaît des anomalies en structure dans les communications ;
- une application à la détection d'intrusion pourrait aussi être intéressante.

4. APPRENTISSAGE ARTIFICIEL ET DONNEES REELLES

Cette partie s'éloigne des tâches de formulation et de développement et présente une analyse autour du problème de la recherche de données dans le domaine de l'apprentissage artificiel. J'ai rencontré cette difficulté au cours de mon stage, car les données disponibles librement sont rares. Elles le sont d'autant plus si l'on cherche des données étiquetées et quasiment inexistante dans le cas de données dynamiques. Deux solutions se présentent en général : l'utilisation de générateurs de données ou récupérer des données réelles.

4.1 GENERATION DE DONNEES

La tâche de génération de données est complexe et représente une tâche de recherche en elle-même. Pour être capable de générer des données avec les caractéristiques voulues, il faut avoir une bonne compréhension générale des mécanismes cachés derrière les données réelles, ce qui est loin d'être évident, notamment dans le cas de données dynamiques comme les réseaux sociaux.

4.1.1 Données formatées

De nombreux articles, comme (Sen, 2007), utilisent des générateurs de données. Leur intérêt principal est de fournir des données dont on connaît les caractéristiques et qui se présentent sous une forme souhaitée.

Comme on le remarque dans l'article, les expériences sur des données de synthèse sont présentées avant celles sur des données réelles. Cette pratique est générale (Zhou, 2004) et on la retrouve aussi dans ce rapport, car la version de la base CORA que j'ai utilisé se rapproche de données générées. Le fond caché derrière cette forme vient de la nature de ces données, de leur formatage, qui forme aussi leur intérêt et de l'utilisation qui en est faite.

Les générateurs de données sont en effet plus utilisés pour réaliser des expériences permettant de confirmer le bon fonctionnement d'un code et d'un modèle et ces expériences sont nécessaires pour présenter son travail à ses pairs.

Il est néanmoins bien sûr possible de réaliser uniquement des expériences sur des données de synthèse pour palier au fait qu'aucune donnée réelle n'était disponible. Mais il faut alors considérer avec recul ces expériences. Car, comme les données générées sont formatées et possèdent des caractéristiques choisies, il est selon moi nécessaire de s'interroger sur l'adaptabilité du modèle à des données à des données réelles qui présenteront des défauts, des incomplétudes, … et cette interrogation est en générale satisfaite par des expériences sur ces données réelles.

4.2 DONNEES REELLES

Les données réelles permettent de tester la robustesse d'un modèle car l'éloignent de la zone de confort que représentent les données de synthèse.

4.2.1 Réelles et propriétaires

Pour le cadre de l'apprentissage sur des données relationnelles, le domaine de l'analyse des réseaux sociaux représente un nouveau champ d'application. Mais aussi l'espérance d'avoir de nouvelles données réelles. Ces dernières sont en général propriétaires et l'entreprise qui les possède doit être d'accord pour en autoriser l'exploitant et la diffusion. Cet accord est en général donné si l'entreprise y trouve un intérêt.

Intérêt pour l'entreprise

La vision traditionnelle de l'entreprise la juge du point de vue de sa rentabilité. Sous cet angle là, on peut comprendre l'intérêt des entreprises pour le domaine : des tâches de

traitement de données comme la classification sont pénibles et longues pour un humain. Utiliser une machine peut donc se révéler un gain de productivité. Cette vision tend maintenant à s'étendre pour prendre en compte l'information comme une ressource primordiale de l'entreprise. Et l'apprentissage automatique y prend tout son sens. On peut par exemple penser aux techniques de fouille de données. Elles permettent d'explorer de très grandes bases de données à la recherche de rapprochements dans les données, ce que ne peuvent pas en général faire des utilisateurs humains. Les méthodes de détection de nouveauté permettent de faire de la veille d'information de manière automatique. Le domaine profite aussi d'une image technologique qui peut plaire au public et donc permettre de vendre des produits. C'est notamment le cas du service Goggles de Google qui est un outil directement issu de la reconnaissance d'images.

Les performances atteintes par le jeune domaine de l'apprentissage artificiel peuvent apparaître prometteuses pour les entreprises. Les tâches couvertes sont diverses : de la classification de documents, à l'aide à la conduite (Mitrovic, 2001), de la transcription de voix à la reconnaissance d'écriture. Et les champs d'application sont donc multiples.

4.2.2 Convergences

Les conférences scientifiques représentent un exemple pertinent de point de convergence entre les entreprises et la communauté scientifique car les deux y apportent quelque chose.

Une méthode de participation originale des entreprises aux conférences et de parrainer un challenger comme par exemple le coupe KDD. En 2009, Orange a parrainé cette coupe (KDD, 2009). On peut essayer de déchiffrer l'intérêt pour cette société. Le premier est une retombée directe : en fournissant des données et un problème qui l'intéresse, Orange obtient des modèles pour le traiter à la pointe de la recherche. Le prix à payer, à savoir fournir et peut-être prétraiter les données est donc faible pour Orange qui, on peut supposer, possède d'énormes bases de données. Le second est une retombée indirecte à court terme : cela tente de renforcer l'image d'Orange comme entreprise tournée vers la recherche et ses liens avec les centres de recherches qui participeront à la conférence.

Cet exemple est illustratif de l'intérêt des entreprises pour le domaine et des motivations derrière ce rapprochement. Le but est de se renseigner sur les techniques à la pointe et de renforcer des liens avec la communauté scientifique.

On y voit aussi le fonctionnement de ce lien. Pour les entreprises, à quelques exceptions près qui disposent de centres de recherche, le développement d'un modèle général ne correspond que rarement à la dynamique de projet qui sous-tend sa stratégie. Au contraire, pour la communauté scientifique, développer un programme résolvant un problème particulier peut apporter de la connaissance, mais généralement moins qu'un problème général. Les deux côtés s'y entendent donc, les entreprises ont besoin que les centres de recherche développent des modèles pour les appliquer, les données et financements des entreprises aident les chercheurs à mettre ces modèles au point. Bien sûr ce lien fonctionne dans la limite que les deux partis y ont plus à gagner qu'à perdre. Si Twitter se permet de laisser de gros volumes de données (Cha M. H., 2010), Facebook ne le fait pas, sûrement car ces données représentent son fond de commerce.

5. CONCLUSION

Ce stage conclut ma double formation d'ingénieur chercheur. Il m'a permis de découvrir le travail au sein de l'univers de la recherche scientifique et d'approfondir mes connaissances dans le domaine de l'apprentissage statistique, ce qui fût la première des difficultés que j'ai rencontrées et que j'ai dépassées grâce à un travail bibliographique. Je poursuivis ce travail afin d'explorer les différents modèles permettant de traiter de la temporalité en essayant de suivre une méthodologie claire pour garder un fil conducteur dans cette recherche et aboutir à la formulation d'un problématique : comment exploiter les relations entre données et commencer à prendre en compte le temps ? Afin de répondre à cette problématique et grâce à l'aide des chercheurs du laboratoire, j'ai mis au point un modèle, de classification par paquet et implémenté des expériences pour le tester.

En plus des problèmes liés à la tâche, comme la recherche de données par exemple, il fût délicat pour moi d'adapter les méthodologies enseignées à TELECOM Bretagne sur l'organisation du travail, notamment lors des projets, à une tâche de recherche. Le but de cette dernière étant plus flou car à plus longue échelle qu'un projet en école d'ingénieur, elle demande plus de rigueur mais est aussi plus intéressante, d'autant qu'il me sera possible de continuer la formalisation de ce modèle ou en proposer de nouveaux le de la thèse CIFRE que je réaliserai avec Thales et le LIP6 qui fait partie d'un projet de l'Agence Nationale de la Recherche sur la détection de fraudes à la carte bancaire sur Internet.

6. ANNEXES

6.1 RESULTATS DE L'EXPERIENCE

Le tableau ci-dessous présente les valeurs numériques de l'expérience réalisée.

Taille	Précision	Rappel	F1	Taille	Précision	Rappel	F1
1	0,613482	0,761979	0,679715	26	0,691813	0,891115	0,778917
2	0,619918	0,77563	0,689087	27	0,736385	0,923294	0,819315
3	0,631059	0,792542	0,702642	28	0,706207	0,903123	0,792618
4	0,636026	0,807845	0,711712	29	0,730974	0,920522	0,81487
5	0,634569	0,8051	0,709734	30	0,720369	0,910927	0,804518
6	0,639568	0,816973	0,717467	31	0,773138	0,943649	0,849926
7	0,643667	0,826798	0,723829	32	0,737461	0,927332	0,821569
8	0,649197	0,833684	0,729964	33	0,786229	0,946898	0,859116
9	0,668106	0,852975	0,749306	34	0,742936	0,918776	0,821553
10	0,663696	0,85624	0,747772	35	0,719225	0,911392	0,803986
11	0,679129	0,868625	0,762277	36	0,72409	0,914741	0,808326
12	0,668873	0,864355	0,754153	37	0,733622	0,920881	0,816654
13	0,691103	0,881027	0,774593	38	0,708889	0,904575	0,794865
14	0,668769	0,86547	0,75451	39	0,723496	0,915941	0,808423
15	0,677147	0,875958	0,763828	40	0,750017	0,93165	0,831025
16	0,684122	0,884204	0,7714	41	0,751063	0,931027	0,831418
17	0,709787	0,901435	0,794213	42	0,747994	0,931709	0,829805
18	0,692765	0,892258	0,779957	43	0,749691	0,92784	0,829306
19	0,719265	0,912943	0,804613	44	0,758084	0,936332	0,837832
20	0,709196	0,907962	0,796364	45	0,802737	0,950912	0,870565
21	0,68684	0,886377	0,773954	46	0,793269	0,953322	0,865962
22	0,684804	0,885812	0,772446	47	0,799771	0,952908	0,869649
23	0,713701	0,907805	0,799135	48	0,786954	0,949241	0,860513
24	0,699862	0,899831	0,787348	49	0,785821	0,94538	0,858247
25	0,723248	0,917498	0,808874	50	0,854805	0,972042	0,909662

6.2 MAIL DU CORPUS ENRON

Voici un exemple de mail de la base de données Enron Flat (Bekkerman) et qui est du même type que ceux du corpus général.

Ce fichier est le fichier beck-s/aec/2.

Message-ID: <6481011.1075855758411.JavaMail.evans@thyme>

Date: Fri, 7 Jul 2000 02:10:00 -0700 (PDT)

From: susan.harrison@enron.com

To: rob.milnthorp@enron.com, sally.beck@enron.com, brent.price@enron.com

Subject: AEC Issues

Cc: bryce.baxter@enron.com

Mime-Version: 1.0

Content-Type: text/plain; charset=us-ascii

Content-Transfer-Encoding: 7bit

Bcc: bryce.baxter@enron.com

X-From: Susan Harrison

X-To: Rob Milnthorp, Sally Beck, Brent A Price

X-cc: Bryce Baxter

X-bcc:

X-Origin: Beck-S

X-FileName: sbeck.nsf

I wanted to provide an update on our progress with AEC. Bryce Baxter and his team members have been in contact with Mike Bennett and Bill Hogue to understand their needs. We have provided the initial information they requested (i.e., one month of data), and are awaiting their approval before completing the analysis on the remaining months. I am confident that we have established clear points of contact within Settlements for Mike and Bill and we will continue to work on solidifying the relationship.

Please let me know if you have any questions.

Susan

6.3 DESCRIPTION DE LA PLATEFORME

La plateforme est un projet découpé pour l'instant en six dossiers : data, dataparser, experience, maths, ml et sqlite3 qui correspondent chacun à des rôles particuliers. L'idée est de proposer un code utile et adaptable à de nombreuses expériences. Le schéma suivant, en pseudo-UML permet de présenter le rôle des objets ainsi que les interactions entre ces derniers.

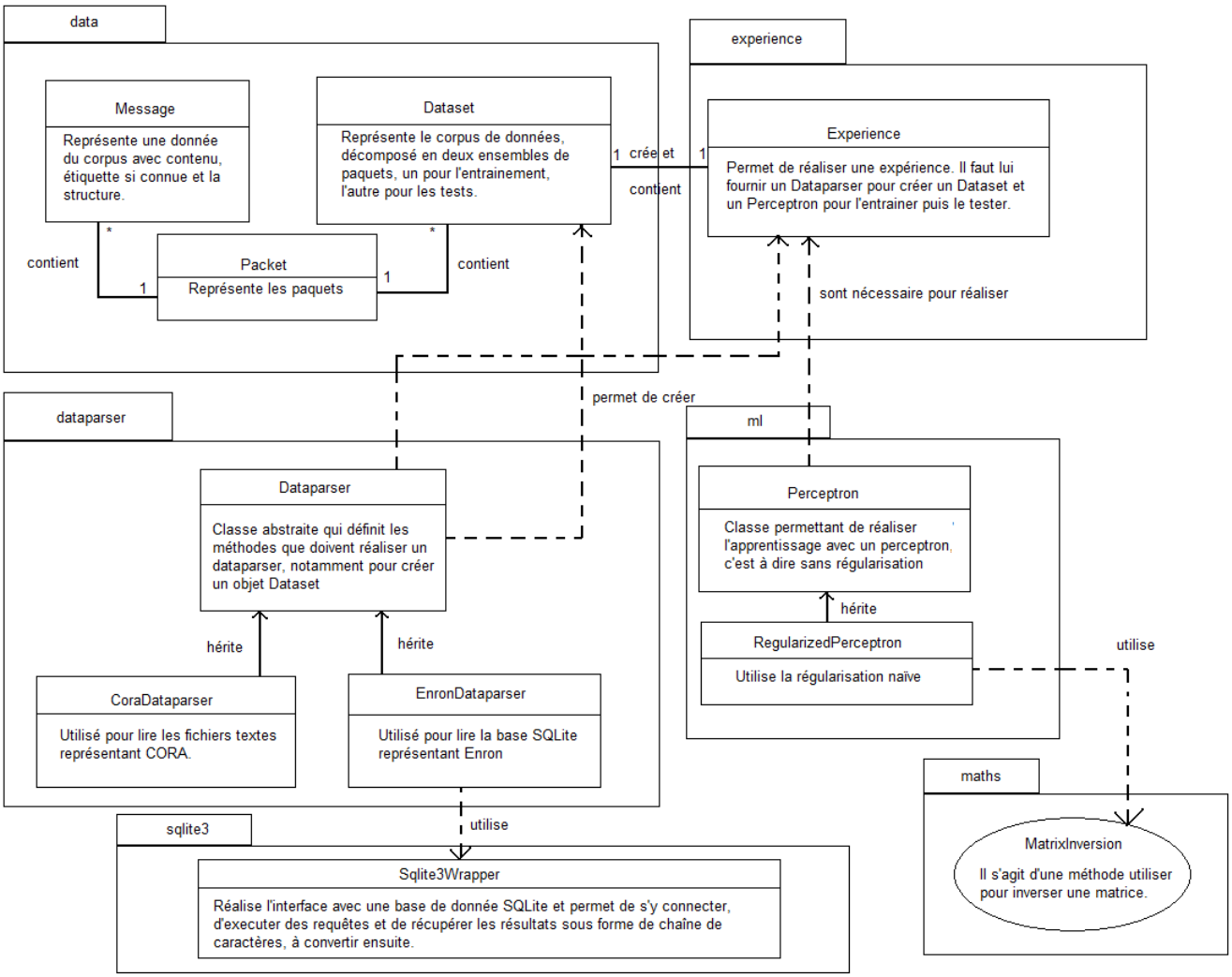

Figure 6 : pseudo-UML de la plateforme développée au cours de mon stage

7. BIBLIOGRAPHIE

Arnold, A. a. (2009). Information Extraction as Link Prediction: Using Curated Citation Networks to Improve Gene Detection. *ICWSM* .

Bekkerman, R. (s.d.). *Ron Bekkerman: Email Classification on Enron Dataset*. Consulté le 08 13, 2010, sur UMass: http://www.cs.umass.edu/~ronb/enron_dataset.html

Boost. (2010). *Boost Basic Linear Algebra*. Consulté le 08 13, 2010, sur Boost C++ Libraries: http://www.boost.org/doc/libs/1_43_0/libs/numeric/ublas/doc/index.htm

Boost. (2010, 08 06). *Boost C++ Libraries*. Consulté le 08 13, 2010, sur Boost C++ Libraries: http://www.boost.org/

Cha, M. H. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICSWM* .

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in twitter: the million follower fallacy. *International Conference on Weblog and Social Media.* Washington DC: ACM.

Chapelle, O. S. (2006). *Semi-supervised Learning.* Boston: MIT Press.

Chi, Y., Song, X., Zhou, D., Hino, K., & Tseng, B. L. (2009). On Evolutionary Spectral Clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (pp. 1-30). ACM.

Cohen, W. W. (2009, 08 11). *Enron Email Dataset*. Consulté le 08 13, 2010, sur Canergie Mellon University - Machine Learning Departement: http://www.cs.cmu.edu/~enron/

Cortes, C. P. (2003). Computational Methods for Dynamic Graphs. *Journal of Computational and Graphical Statistics, Vol. 12, No. 4, Statistical Analysis of Massive Data Streams* , 950-970.

De Choudhury, M. L. (2010). How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? *ICWSM* .

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring Social Network Structure using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, (pp. 15274-15278).

Fu, W. a. (2009). Dynamic mixed membership blockmodel for evolving networks. *Proceedings of the 26th Annual International Conference on Machine Learning* , 329-336.

KDD. (2009). *KDD Cup 2009*. Consulté le 08 13, 2010, sur KDD Cup 2009: http://www.kddcup-orange.com/

Kempe, D., Kleinberg, J., & Kumar, A. (2000). Connectivity and Inference Problems for Temporal Networks. *Proceedings of the Thirty-Second Annual ACM Symposium on theory of Computing* (pp. 504 - 513). Portland, Oregon, United States: ACM.

Lin, Y. a. (2008). Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. *Proceeding of the 17th international conference on World Wide Web* , 685-694.

LINQS.UMD . (s.d.). *Projects @ LINQS.UMD*. Consulté le 08 13, 2010, sur Lise's Inquisive Students - Machine Learning Research Group @ UMD: http://www.cs.umd.edu/~sen/lbc-proj/LBC.html

LIP6. (2010). *Département «Calcul Scientifique»*. Consulté le 08 05, 2010, sur Laboratoire d'Informatique de Paris 6: http://www.lip6.fr/recherche/team.php?id=100

LIP6. (2010). *Département «DEcision, Systèmes Intelligents Recherche opérationnelle»*. Consulté le 08 05, 2010, sur http://www.lip6.fr/recherche/team.php?id=300

LIP6. (2010). *Département «Données et Apprentissage Artificiel»*. Consulté le 08 05, 2010, sur http://www.lip6.fr/recherche/team.php?id=500

LIP6. (2010). *Département «Réseaux et Systèmes Répartis»*. Consulté le 08 05, 2010, sur http://www.lip6.fr/recherche/team.php?id=700

LIP6. (2010). *Département «Systèmes Embarqués sur Puce»*. Consulté le 08 05, 2010, sur http://www.lip6.fr/recherche/team.php?id=900

Masud, M. M. (2010). Classification and Novel Class Detection of Data Streams in A Dynamic Feature Space. *PKDD* .

McCulloh, I. a. (2008). *Social Network Change Detection.* Carnegie Mellon University Technical Report, CMU-CS-08-116.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Science.

Mitrovic, D. (2001). Machine Learning for Car Navigation. *Lecture Notes in Computer Science* .

Parikh, N., & Sundaresan, N. (2008). Scalable and near real-time burst detection from eCommerce queries. *Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (pp. 24 - 27). Las Vegas: ACM.

Peter Bodik, W. H. (2004, 06 02). *Intel Lab Data*. Consulté le 08 13, 2010, sur Intel Lab Data: http://db.csail.mit.edu/labdata/labdata.html

Sadikov, E. P. (2009). *Blogs as Predictors of Movie Success.* Stanford Labs.

Sen, P. a. (2007). Link-based classification. *University of Maryland Technical Report CS-TR-4858* .

Sharan, U. a. (2008). Temporal-Relational Classifiers for Prediction in Evolving Domains. *ICDM* .

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* , 11–21.

SQLite. (2010). *SQLite Home Page*. Consulté le 08 13, 2010, sur SQLite: http://www.sqlite.org/

Tang, J. a. (2009). Temporal distance metrics for social network analysis. *Proceedings of the 2nd ACM workshop on Online social networks* , 31-36.

Wikimedia. (2010, Avril 08). *Précision et rappel - Wikipédia*. Consulté le 08 13, 2010, sur Wikipédia: http://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel

Zhou, D. B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems 16 : Proceedings of the 2003 Conference* , 595-602.

8. GLOSSAIRE

Terme :	Définition :
LIP6	Laboratoire d'Informatique de Paris 6, 4 place Jussieu – 75005 Paris
UPMC	Université Pierre et Marie Curie.
CNRS	Centre National de la Recherche Scientifique.
MALIRE	MAchine Learning and Information REtrieval, équipe du département DAPA du LIP6 spécialisé en apprentissage.
DAPA	Données et APprentissage Artificiel, département du LIP6.
CORA	Moteur de recherche pour articles de recherche. Le nom est utilisé pour désigner une base d'articles utilisée couramment en classification.
Classification	Tâche consistant à trouver la classe de données
Classe	Nom que l'on peut attacher à des données pour résumer un ensemble de propriété communes entre elles
Etiquette	Cf. classe, utilisé dans le domaine de la classification comme traduction de *label*
Caractéristiques	Valeurs qui sont désignées pour représenter une donnée.
Oracle	Nom utilisé pour désigner l'ensemble de règles permettant de connaître la classe d'un objet.
Cluster	Regroupement ou partition à l'intérieur des données

Clustering	Recherche de clusters sur un ensemble de données.
Graphe	Représentation mathématique et graphique d'un ensemble de données liées entre-elles.
Précision	Score permettant de quantifier si un algorithme de classification délimite correctement les classes.
Rappel	Score permettant de quantifier si un algorithme de classification reconnaît correctement les classes.
F-score	Score permettant de résumer précision et rappel.
Bayes (Thomas)	Mathématicien britannique.
Théorème de Karush-Kuhn-Tucker	(ou conditions de Karush-Kuhn-Tucker) permet de reformuler des problèmes d'optimisation pour les résoudre plus facilement.
Echantillonnage de Gibbs	Algorithme d'échantillonnage permettant une approximation d'une probabilité conjointe.
Variables latentes	Variables non observés mais interprétées qui permettent parfois de caractériser un problème.
Enron	Entreprise énergétique texane ayant fait faillite en 2001.
MIME	Multipurpose Internet Mail Extensions. Format représentant les courriels.
SQL	Structured Query Language. Langage normalisé utilisé pour interroger des bases de données.
SQLite	Une implémentation du langage SQL
Boost	Librairie C++

www.telecom-bretagne.eu

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France
Tél. : + 33 (0)2 29 00 11 11
Fax : + 33 (0)2 29 00 10 00

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
France
Tél. : + 33 (0)2 99 12 70 00
Fax : + 33 (0)2 99 12 70 19

Campus de Toulouse
10, avenue Edouard Belin
BP 44004
31028 Toulouse Cedex 04
France
Tél. : +33 (0)5 61 33 83 65
Fax : +33 (0)5 61 33 83 75