Stéphane Girard,
Directeur de recherche,
Inria Grenoble Rhône-Alpes,
équipe-projet MISTIS,
Inovallée, 655, av. de l'Europe,
Montbonnot, 38334 Saint-Ismier cedex.
Email : Stephane.Girard@inria.fr

Le 19 octobre 2016.

Rapport sur la thèse présentée par Nicolas Goix

Le manuscrit proposé par Nicolas Goix en vue d'obtenir le doctorat spécialité "Signal et Images" de Télécom ParisTech s'intitule

"Apprentissage automatique et extrêmes pour la détection d'anomalies".

Le manuscrit est rédigé en anglais, il est formé de quatre parties, hors introduction et conclusion. La première partie est composée de préliminaires dédiés aux inégalités de concentration, théorie des valeurs extrêmes, et algorithmes classiques de détection d'anomalies. Les trois parties suivantes présentent les contributions du candidat : détection d'anomalies via l'introduction de la courbe *Excess-Mass* (Partie 2), estimation de la fonction de dépendance de queue et représentation parcimonieuse des extrêmes multivariés (Partie 3), évaluation des performances des algorithmes de détection d'anomalies et introduction d'une méthode de forêts aléatoires à une classe (Partie 4).

L'introduction présente le contexte et les contributions de la thèse. Pour résumer rapidement, la détection d'anomalies supervisée est un problème de classification binaire où l'une des deux classes est sous-représentée dans l'ensemble d'apprentissage. La détection d'anomalies non-supervisée est un problème de détection de points atypiques, c'est-à-dire en dehors du modèle. Le cadre semi-supervisé correspond à la situation où seuls les individus "normaux" sont observés, situation parfois appelée classification à une seule classe. Les résultats obtenus dans la thèse de Nicolas Goix sont décrits de façon pédagogique. Ils sont résumés de manière concise et illustrés judicieusement, les points saillants sont bien mis en évidence.

La première partie du manuscrit est structurée en trois chapitres, le premier d'entre eux étant dédié aux inégalités de concentration. Nicolas Goix rappelle deux inégalités de concentration fondamentales (l'une pour les variables bornées, l'autre basée sur un contrôle de la variance) ainsi que leurs preuves et montre comment les inégalités classiques (Hoeffding, Bernstein) peuvent en être déduites. Le point essentiel de ce chapitre est de montrer com-

ment l'inégalité de Vapnik-Chervonenkis, à la base de nombreux travaux en apprentissage statistique, peut être déduite des inégalités de concentration et surtout affinée pour des zones de faible probabilité. Il s'agit donc d'un des résultats majeurs de la thèse de Nicolas Goix. Il est par exemple exploité dans la Partie 3 du manuscrit mais il ouvre la porte à beaucoup d'autres applications en statistique des valeurs extrêmes. Dans ce domaine, les résultats sont en effet souvent asymptotiques, et ce nouvel outil permettrait d'établir des nouvelles bornes non-asymptotiques. Le second chapitre est de nature bibliographique, il rappelle les points clefs de la théorie des valeurs extrêmes univariée et multivariée. Le troisième chapitre dresse un état de l'art sur la détection d'anomalies. Trois méthodes classiques sont mises en évidence : *One-class SVM* (Séparateurs à Vaste Marge à une classe) qui estiment les lignes de niveaux de la densité, *Local Outlier Factor* basé sur l'algorithme des plus proches voisins et *Isolation Forest* adapté de l'algorithme des forêts aléatoires. Les deux dernières méthodes ont été implémentées par Nicolas Goix en Python au sein de la bibliothèque *scikit-learn*. Il aurait été intéressant de quantifier le volume de travail fourni en termes de nombre de lignes de code par exemple. L'utilisation des fonctions développées est illustrée par des exemples de commandes Python accompagnés de sorties graphiques.

La seconde partie du manuscrit est dédiée à l'introduction d'un critère de performance pour la détection d'anomalies appelé courbe *Excess-Mass*. Ce travail a fait l'objet d'une communication à la conférence AISTATS 2015. La détection d'anomalies repose sur l'estimation de scores qui, idéalement, sont des fonctions croissantes de la densité, et mesurent le degré d'anomalie d'une observation. Afin de comparer plusieurs scores, et potentiellement trouver le score optimal, des critères ont été introduits récemment dans la littérature tels que la courbe *Mass-Volume* qui peut être interprétée comme une extension de la courbe ROC au cas non-supervisé. Nicolas Goix propose de pallier certaines limitations de ce critère en définissant la courbe *Excess-Mass* qui induit un ordre partiel sur les fonctions scores. Un procédé d'optimisation du critère *Excess-Mass* est également introduit afin d'estimer la fonction score optimale. Une borne non-asymptotique est construite pour contrôler l'erreur entre la courbe *Excess-Mass* optimale et son estimation. Le comportement de cette erreur est illustré sur simulations.

La Partie 3 comporte deux chapitres consacrés à des travaux réalisés dans le cadre de la statistique des valeurs extrêmes multivariées. L'estimation de la fonction de dépendance de queue (*stable tail dependence function*), qui est une caractérisation possible de la dépendance des marginales dans la queue d'une loi, est étudiée Chapitre 6. Ce travail a fait l'objet d'une communication à la conférence COLT 2015. En se basant sur les résultats de concentration établis dans la Partie 1, Nicolas Goix établit une borne non-asymptotique sur l'estimateur empirique de la fonction de dépendance de queue. Ce résultat complète les études asymptotiques récentes, telles que la normalité asymptotique prouvée par Einmahl *et. al.*, 2012. Ici, le candidat a mis l'accent sur l'étude du terme de variance en précisant que le terme de biais peut être analysé sous des hypothèses de second ordre, ce qui est un choix raisonnable. Une autre application des inégalités de concentration est présentée en classification dans des zones de faible probabilité. Ces

deux exemples montrent bien le potentiel des inégalités de concentration établies par le candidat dans des zones de faible probabilité pour comprendre le comportement des estimateurs classiques. L'utilisation pour construire de nouveaux estimateurs ou tests dans de telles région ne semble par contre pas immédiate. Une caractérisation alternative à la fonction de dépendance de queue pour la loi des extrêmes multivariées est la mesure angulaire. Nicolas Goix propose dans le Chapitre 7 une modélisation et une estimation parcimonieuse de cette fonction. A ma connaissance, il s'agit de la première tentative de ce type pour la modélisation des extrêmes en grande dimension. Ce travail est actuellement soumis pour publication et a été présenté à la conférence AISTAT 2016 ainsi qu'à un workshop satellite de NIPS 2015. L'idée est d'estimer empiriquement la masse répartie par la mesure angulaire au voisinage de sous-cônes de l'espace de dimension d. Les performances non-asymptotiques de l'estimateur sont établies avec des techniques similaires à celles du chapitre précédent. Les régions de faible masse étant ignorées, cela conduit à une représentation parcimonieuse de la mesure angulaire et permet l'introduction d'une fonction score pour la détection d'anomalies. L'algorithme résultant est baptisé DAMEX et sa complexité est selon l'auteur de l'ordre de $dn \log n$ où n désigne la taille de l'échantillon. Ce point aurait pu être davantage détaillé, j'aurais aimé voir souligné pourquoi la complexité n'est pas proportionnelle à 2^d, le nombre de sous-cônes. L'estimation du support de la mesure angulaire est illustrée sur des données simulées et des données réelles. Sur ces dernières, il apparaît que la modélisation choisie permet de capturer de manière efficace la dépendance dans les extrêmes grâce à un petit nombre de sous-cônes de dimensions modérées. L'algorithme DAMEX est comparé avec soin à *Isolation Forest* sur cinq jeux de données classiques dans le domaine de la détection d'anomalies. Ses performances sont sensiblement meilleures sur les régions extrêmes.

La Partie 4 présente deux contributions supplémentaires à la détection d'anomalies. Comme le signale l'auteur, comparés aux travaux précédents, ceux-ci sont de nature plus expérimentale. Le Chapitre 8 aborde la question de la comparaison des performances de méthodes de détection d'anomalies, il a fait l'objet d'une communication à un workshop satellite d'ICML 2016. Le principe de comparaison est basé sur le calcul de l'aire sous les courbes *Excess-Mass* et *Mass-Volume* introduites dans la seconde partie. Afin d'obtenir un procédé utilisable en grande dimension, le candidat propose de réaliser un échantillonnage aléatoire parmi les variables, suivi d'une agrégation par moyenne des résultats partiels obtenus. La procédure est testée sur douze jeux de données standards. Dans le cas supervisé, elle peut être comparée avec deux critères dédiés à cette situation : les aires sous les courbes ROC et précision-rappel. Il apparaît que les quatre critères donnent des résultats comparables. Dans le cadre non-supervisé, les deux critères basés sur les courbes *Excess-Mass* et *Mass-Volume* semblent donc des extensions pertinentes des critères classiques basés sur les courbes ROC et précision-rappel. Le Chapitre 9 propose une adaptation de la méthode des forêts aléatoires à la détection d'anomalies semi-supervisée ou classification à une seule classe. La tâche principale est d'adapter le critère de découpage initialement prévu pour le cas de deux classes au cas d'une seule classe. Nicolas Goix introduit deux approches basées sur deux approximations différentes

de la densité de points aberrants : une approximation uniforme (approche naïve) et une approximation uniforme par morceaux (approche adaptative). La méthode ainsi obtenue est comparée sur les mêmes douze jeux de données qu'au Chapitre 8 à sept algorithmes classiques de détection d'anomalies dont *One-class SVM*, *Local Outlier Factor* et *Isolation Forest* déjà mentionnés ici. Les performances sont évaluées en termes d'aires sous les courbes ROC et précision-rappel. La méthode proposée offre en moyenne les meilleures performances avec un temps de calcul très compétitif.

Le manuscrit se termine par un chapitre de conclusion et perspectives. Les différentes contributions de la thèse y sont résumées de façon claire et synthétique. Les perspectives proposées sont raisonnables.

En conclusion, les recherches menées dans cette thèse couvrent deux domaines de la statistique rarement réunis jusqu'à présent : l'apprentissage statistique et la statistique des valeurs extrêmes. En cela, les travaux de Nicolas Goix sont très originaux et ouvrent la voie à une étude non-asymptotique des événements extrêmes. Les applications en détection d'anomalies développées dans la thèse en attestent. Par ailleurs, le manuscrit présente un bel équilibre entre les aspects appliqués et théoriques de la recherche en statistique. Il illustre la capacité de Nicolas Goix à proposer des méthodes originales, à établir leurs propriétés théoriques ainsi qu'à les implémenter informatiquement pour les confronter avec les techniques existantes sur données réelles.

Pour ces raisons, je suis très favorable à ce que cette thèse soit soutenue en l'état.

Stéphane Girard