# Apprentissage Automatique et Extrêmes pour la Détection d'Anomalies

**Nicolas Goix**

**RESUME.** La détection d'anomalies est tout d'abord une étape utile de pré-traitement des données pour entraîner un algorithme d'apprentissage statistique. C'est aussi une composante importante d'une grande variété d'applications concrètes, allant de la finance, de l'assurance à la biologie computationnelle en passant par la santé, les télécommunications ou les sciences environnementales. La détection d'anomalies est aussi de plus en plus utile au monde contemporain, où il est nécessaire de surveiller et de diagnostiquer un nombre croissant de systèmes autonomes. La recherche en détection d'anomalies inclut la création d'algorithmes efficaces accompagnée d'une étude théorique, mais pose aussi la question de l'évaluation de tels algorithmes, particulièrement lorsque l'on ne dispose pas de données labellisées – comme dans une multitude de contextes industriels. En d'autres termes, l'élaboration du modèle et son étude théorique, mais aussi la sélection du modèle.

Dans cette thèse, nous abordons ces deux aspects. En pratique, un algorithme de détection d'anomalies retourne une *fonction de score* à valeurs réelles définie sur l'espace des données de manière à quantifier l'anormalité des observations. Tout d'abord, nous introduisons un critère alternatif au critère masse-volume existant, pour mesurer les performances d'une fonction de score. Ce critère, appelé *critère d'excès de masse*, à pour but la construction de fonctions de score *via* la minimisation du risque empirique.

La seconde partie de ce travail porte sur les régions *extrêmes*, qui sont d'un intérêt particulier en détection d'anomalies. En particulier, des outils probabilistes issues de la théorie des valeurs extrêmes (multivariées), comme la STDF (stable tail dependence function) et la mesure angulaire, peuvent être combinés avec une approche plus classique de détection d'anomalies afin de gagner en précision sur ces régions extrêmes. Des bornes non-asymptotiques sont établies pour l'estimation de la STDF, cette dernière caractérisant la structure de dépendance dans les extrêmes. Une méthode statistique produisant une représentation (possiblement parcimonieuse) de la structure de dépendance est ensuite dérivée de l'estimation non-paramétrique de la mesure angulaire restreinte à un ensemble représentatif de directions. Cette représentation peut être utilisée pour produire une fonction de score précise sur les régions extrêmes. Des bornes non-asymptotiques attestant de la qualité de l'estimation sont établies.

La dernière partie de ce travail est essentiellement de nature heuristique. D'un point de vue sélection de modèle, nous étudions empiriquement l'usage des courbes masse-volume et d'excès de masse (en l'absence de données labellisées). Comme ces courbes ne peuvent généralement pas être estimées avec qualité en grande dimension, une méthode basée sur le sous-échantillonnage de variables est aussi développée et testée, étendant ainsi l'usage de ces deux critères à des jeux de données de grande dimension. Du point de vue élaboration de modèle, un algorithme efficace basé sur les forêts aléatoires et produisant des fonctions de score précises est proposé. Cet algorithme se base sur une extension naturelle des critères de séparation standards au cas où une seule classe est observée, et donne en pratique des performances remarquables selon une étude comparative incluant une multitude d'algorithmes de détection d'anomalies utilisés dans l'industrie.

**Mots-Clefs:** Détection d'anomalies, extrêmes multivariés, forêts aléatoires, sélection de modèles non-supervisé.

**ABSTRACT.** Anomaly detection is not only a useful preprocessing step for training machine learning algorithms. It is also a crucial component of many real-world applications, from various fields like finance, insurance, telecommunication, computational biology, health or environmental sciences. Anomaly detection is also more and more relevant in the modern world, as an increasing number of autonomous systems need to be monitored and diagnosed. Important research areas in anomaly detection include the design of efficient algorithms and their theoretical study but also the evaluation of such algorithms, in particular when no labeled data is available – as in lots of industrial setups. In other words, model design and study, and model selection.

In this thesis, we focus on both of these aspects. In practice, anomaly detection algorithms output a real valued *scoring function* on the feature space so as to quantify to which extent observations should be considered as abnormal. We first propose a criterion for measuring the performance of scoring functions, alternative to the existing *mass-volume curve*. This criterion, referred to as the *excess-mass curve*, aims at building scoring functions *via* empirical risk minimization.

The second part of this work focuses on *extreme* regions, which are of particular interest in anomaly detection. In particular, probabilistic tools borrowed from (multivariate) extreme value theory, such as the stable tail dependence function (STDF) and the angular measure, can be combined with classical anomaly detection approaches to gain in accuracy on such extreme regions. We provide non-asymptotic bounds for the estimation of the STDF, which characterizes the extreme dependence structure. A statistical method that produces a (possibly sparse) representation of the extreme dependence structure is then derived from a non-parametric estimation of the angular measure on representative sets of directions. This representation can be used to produce a scoring function accurate on extremes regions. Non-asymptotic bounds to assess the accuracy of the estimation procedure are established.

The last part of this work is essentially of heuristic nature. From the model selection viewpoint, an empirical study for the use of excess-mass and mass-volume curves as evaluation criteria (in the absence of labeled data) is derived. As these curves generally cannot be well estimated in large dimension, a methodology based on feature sub-sampling and aggregating is also described and tested, extending the use of these criteria to high-dimensional datasets. From the model design viewpoint, an efficient algorithm based on random forests producing accurate scoring functions is proposed. It builds on a natural extension of standard splitting criteria to the one-class setting, and competes well according to an extensive benchmark which includes many state-of-the-art anomaly detection algorithms, commonly used in industrial setups.

**Keywords:** Anomaly detection, multivariate extremes, random forests, unsupervised model selection.