

# Outils Statistiques pour l'Évaluation des Performances en Classification et Apprentissage en présence de données entachées d'erreurs

## Statistical Performance Evaluation Tools for Classification and Machine Learning with Erroneous Data

*Proposition de sujet de thèse Fédération Charles Hermite.*

Loria – IECL

### Encadrants

- [Angelo Efoevi Koudou](#) – MCF HDR – [IECL](#) (équipe Probabilités et Statistique)
- [Bart Lamiroy](#) – MCF HDR – [Loria](#) (équipe Synalp)

### Contexte du sujet

Ce sujet de thèse est la prolongation du PEPS CNRS 2016 « *Perfaclastique* » entre le Loria et l'IECL et le projet Mastodons « *Apprentistique* » 2017 du CNRS entre le Loria, l'IECL et l'INRA de Toulouse.

### Mots clés

Algorithmes de classification ; évaluation de performance ; divergence de Kullback-Leibler ; indice de Rand ; régression ; apprentissage statistique.

### Keywords

Classification ; performance evaluation, Kullback-Leibler divergence ; Rand Index ; regression ; machine learning

### Sujet

Le but de cette thèse est de contribuer à l'état de l'art en évaluation de performances dans des problèmes de classification (notamment en perception artificielle) en se positionnant en rupture par rapport aux consensus établis. Les méthodes d'apprentissage et de classification actuelles dépendent très fortement de grandes masses de données annotées pour fonctionner. Le bouleversement de l'état de l'art, notamment provoqué par les méthodes d'apprentissage profond, nécessite des approches d'évaluation des performances adaptées. Les hypothèses traditionnelles sur les données de référence pour mesurer les performances se trouvent fortement affaiblies du fait de la quantité des données nécessaire pour les faire fonctionner. Il devient alors impossible de présupposer que les données d'apprentissage et/ou d'évaluation soient exemptes d'erreurs ou de bruit, induisant ainsi des imprécisions sur les évaluations et les comparaisons entre expérimentations.

Dans ce projet, nous visons à établir à la fois les formalismes mathématiques et de protocoles expérimentaux qui permettront d'exprimer des niveaux de confiance et des métriques statistiques, pour prendre en compte l'incertitude sur les données dans l'évaluation de méthodes de classification et d'apprentissage.

Nous proposons de revisiter l'ensemble du processus en étudiant et en développant des outils statistiques permettant d'exprimer une « confiance » dans des mesures de classement issues de campagnes d'évaluation ou de *benchmarking*.

La question à laquelle on cherchera à répondre est la suivante :

*Étant donnée la réponse de  $n$  algorithmes sur un ensemble de données de référence, quelle est la confiance que l'on peut accorder au classement qui en résulte, sachant que le taux d'erreur des données*

de référence est inférieur à  $\epsilon$ . Ou à partir de quel taux d'erreur sur les données de référence peut-on considérer, avec un taux de certitude de  $\tau$ , que le classement obtenu sera mis en défaut. De façon duale, on peut également, non pas exprimer une confiance dans les classifieurs, mais dans les données expérimentales.

Plusieurs formulations probabilistes de cette question sont possibles. Par exemple, *en considérant les données comme des réalisations d'un vecteur aléatoire (dont la loi pourra appartenir à un modèle paramétrique donné), on étudiera la loi du vecteur constitué des réponses des  $n$  algorithmes en tant que fonction de ce vecteur aléatoire, ce qui permettra de calculer la probabilité d'avoir un classement donné de ces réponses, et de considérer que le classement n'est pas fiable si cette probabilité est jugée trop faible. En supposant que la distribution des erreurs appartient à un modèle paramétrique, des outils de statistique bayésienne pourront être utilisés pour étudier la distribution a posteriori des paramètres au vu des réponses des algorithmes.*

## Topic

Classical performance evaluation metrics used in the domains of Classification and Machine Learning are based on the assumption that the reference data used for validation and comparison are not error-free. Recent work has shown that this hypothesis is almost never guaranteed.

In order to evaluate and compare different classification methods objectively and with verifiable levels of confidence, we have started to explore empirical statistical techniques which are promising and which make it possible to avoid of the constraint of perfect data. These techniques now need to be formally and theoretically validated on the one hand and extended to classifiers that are more generic. Our experiments validated the approach on binary classifiers; they should be generalized to classifiers involving any number of classes.

This project aims to revisit the whole performance evaluation process by studying and developing statistical tools expressing a 'confidence' in classification measures resulting from evaluation or benchmarking campaigns.

The question to be answered is the following: *given the response of  $n$  algorithms to a set of reference data, what confidence can be attached to the resulting ranking, given an estimated error rate of less than  $\epsilon$ ? Or, conversely, from what error rate on the reference data can we consider, with a confidence rate of  $\tau$  that the obtained classification cannot be guaranteed?* These questions can be expressed with several probabilistic formulations. For example, considering the data as realizations of a random variable (whose law may belong to a given parametric model), we can study the responses of the  $n$  algorithms as a function of this random variable. This will make it possible to compute the probability of having a given ranking of these answers and to test if the ranking is reliable. Assuming that the error distribution belongs to a parametric model, Bayesian statistical tools can be used to study the posterior distribution of the parameters in the light of the responses of the algorithms.

## Références

1. Lamiroy, B., Sun, T.: « Computing Precision and Recall with Missing or Uncertain Ground Truth ». In: Kwon, Y.B., Ogier, J.M. (eds.) Graphics Recognition. New Trends and Challenges. 9th International Workshop, GREC 2011, Seoul, Korea, September 15-16, 2011, Revised Selected Papers, Lecture Notes in Computer Science, vol. 7423, pp. 149–162. Springer (Feb 2013),
2. Lamiroy, B., Pierrot, P. « Statistical Performance Metrics for Use with Imprecise Ground-Truth » In: Lamiroy, B., Dueire Lins, R. (eds.) Graphics Recognition. Current Trends and Challenges, 11th IAPR International Workshop, GREC 2015. Revised Selected Papers, Lecture Notes in Computer Science, vol. 9657, Springer 2017
3. Rand, W. M. « *Objective criteria for the evaluation of clustering methods* », [\*Journal of the American Statistical Association\*](#), American Statistical Association, vol. 66, n° 336, 1971, p. 846–850.

4. Robert, C. (2001). *The Bayesian Choice*. Springer Verlag, collection Springer Texts in Statistics.

## Environnement

Ce travail est co-encadré entre deux équipes de recherche, l'une spécialisée dans la classification et l'apprentissage, l'autre dans les statistiques. Les candidat.e.s pourront, sans distinction, mettre en valeur un profil plus *informatique* ou *mathématique* selon leurs compétences. Il est attendu que les candidat.e.s investissent ensuite le champ scientifique leur correspondant le mieux, tout en gardant une ouverture et une interaction suffisante avec l'autre.

Le travail effectué dans cette thèse, s'inscrit dans le projet CNRS Mastodons 2017 « *Apprentistrique* » et est également financé par la Fédération Charles Hermite (FR 3198).

Les candidat.e.s potentiel.le.s doivent prendre contact avec les encadrants en envoyant un CV et lettre de motivation, et obligatoirement faire acte de candidature via <http://www.adum.fr/as/ed/page.pl?site=IAEM&page=candidater>.

This work being co-supervised by two research teams (one specialized in classification and learning, the other in statistics), the candidate is required to develop either but without distinction, a more Computer science profile, either a more Applied Mathematics profile, or both, depending on their skills. It is expected that candidates will then invest the most appropriate scientific field, while maintaining sufficient openness and interaction with the other.

In general, candidates will have to have a strong Computer Science and Mathematical literacy and a exhibit a high level of curiosity in the fields of classification and machine learning.

Potential candidates should send a resume and motivation letter and apply online through <http://www.adum.fr/as/ed/page.pl?site=IAEM&page=candidater>.