Évaluation de critères de sélection de noyaux pour la régression Ridge à noyau dans un contexte de petits jeux de données

Frédérick Fabre Ferber, Dominique Gay, Jean-Christophe Soulié, Jean Diatta, Odalric-Ambrym Maillard

Évaluation de critères de sélection de noyaux pour la régression Ridge à noyau dans un contexte de petits jeux de données

Frederick Fabre Ferber[*,**], Dominique Gay[*], Jean-Christophe Soulie[**],
Jean Diatta[*], Odalric-Ambrym Maillard[***]

[*]LIM-EA2525, Université de La Réunion, France
frederick.fabre, jean.diatta, dominique.gay@univ-reunion.fr,

[**]CIRAD, UPR Recyclage et Risque, F-97743 Saint-Denis, Réunion, France
Recyclage et Risque, Univ Montpellier, CIRAD, Montpellier, France
frederick.fabre-ferber, jean-christophe.soulie@cirad.fr

[***]INRIA Lille - Nord Europe, SCOOL, France
odalric.maillard@inria.fr

1 Introduction

Lorsque l'on travaille avec des données réalistes et complexes, comme en agronomie ou en écologie (Hunt et al., 2001), il est plus compliqué de collecter des données automatiquement et efficacement, en raison de l'expertise requise pour étiqueter les données correctement. Il en résulte peu d'observations, ce qui complique la tâche des algorithmes d'apprentissage. Dans de tels cas, les méthodes à noyau (Hofmann et al., 2008) peuvent être une bonne solution, par rapport à d'autres familles d'algorithmes. La sélection du noyau le plus approprié pour un critère donné est essentielle pour garantir de bonne performances prédictives. Il est donc nécessaire de disposer d'une mesure (critère) permettant de dire si un noyau est meilleur qu'un autre pour une tâche donnée. Nous proposons d'évaluer divers critères de sélection de noyaux issus de la littérature, pour la régression à noyau ridge dans un contexte de petits ensembles de données synthétiques et issues du réel.

2 Expérimentations

Nous évaluons en terme de performance à travers la MSE (Mean Squared Error) 6 critères de sélection pour 6 jeux de données différents (Housing, Air Quality, Agronomy, Energy Efficiency, Forest Fire, Kernel Based Dataset). Nous utilisons comme algorithme la Regression Ridge à Noyau avec le noyau RBF $k(x, x') = \mathbf{exp}(-\frac{||x-x'||}{2\sigma^2})$, où le paramètre σ prend ses valauers dans $[0.02, 0.04, \ldots, 1]$.

Nous avons entrepris une analyse approfondie de divers critères de sélection de paramètres de noyau dans le contexte de l'apprentissage automatique, en mettant l'accent sur l'évaluation basée sur l'erreur quadratique moyenne (MSE) sur plusieurs ensembles de données. Nos résultats mettent en lumière des tendances intéressantes. La plupart des critères de sélection ont des performances comparables, suggérant qu'ils sont tous efficaces pour choisir un "bon" noyau, sauf pour le KTA (Kernel Target Alignment), pour lequel les performances prédictives sont moindres (Table 1). Nous avons également exploré la capacité des critères à trouver le noyau correct lorsque la nature du signal est connue, ainsi que leur résilience à différents bruits dans les données d'apprentissage. Nos expériences ont montré que, dans ces scénarios, tous les critères, à l'exception du KTA, se sont révélés équivalents. L'évolution de l'erreur vers la valeur du bruit avec l'augmentation du nombre d'observations suggère que les noyaux sélectionnés par les critères sont adaptés au modèle.

Datasets	AIC	BIC	RMSE	CONF	EIG	KTA
Housing	**2.14 ± 0.05**	**2.14 ± 0.05**	**2.14 ± 0.05**	**2.14 ± 0.05**	**2.14 ± 0.05**	5.51 ± 0.04
Air Quality	781 ± 77	781 ± 77	781 ± 77	781 ± 77	781 ± 77	781 ± 77
Agronomy	**1.32 ± 0.07**	**1.32 ± 0.07**	**1.32 ± 0.07**	**1.32 ± 0.07**	**1.32 ± 0.07**	1.53 ± 0.07
Energy Eff.	**11.5 ± 0.1**	**11.5 ± 0.1**	**11.5 ± 0.1**	**11.5 ± 0.1**	**11.5 ± 0.1**	94.02 ± 12.05
Forest Fire	**4540 ± 2127**	**4540 ± 2127**	**4540 ± 2127**	4763 ± 1814	4763 ± 1814	**4540 ± 2127**
Kernel Data	**0.6 ± 0.11**	**0.6 ± 0.11**	**0.6 ± 0.11**	**0.6 ± 0.11**	**0.6 ± 0.11**	35.3 ± 4.94

TAB. 1 – *Comparaison des performances en terme de MSE pour les différents jeux de données et critères. Les meilleurs performances sont soulignées et en gras, excepté pour AirQuality où les performances sont équivalentes pour tous les critères.*

3 Conclusion

En conclusion, la sélection d'un paramètre de noyau approprié reste un défi pour l'apprentissage automatique, dans un contexte de petits jeux de données. Nous avons montré que tous les critères (hormis le KTA) étaient équivalents ce qui laisse la possibilité de choisir entre des critères offrant une garantie explicite de l'erreur de prédiction et des critères plus simples, moins exigeants en termes de temps de calcul.

Références

Hofmann, T., B. Schölkopf, et A. J. Smola (2008). Kernel methods in machine learning. *The Annals of Statistics 36*(3), 1171–1220.

Hunt, L., J. White, et G. Hoogenboom (2001). Agronomic data : advances in documentation and protocols for exchange and use. *Agricultural Systems 70*(2-3), 477–492.