



إستخدام بعض الطرائق في اختيار المتغيرات ذات الابعاد العالية في السلاسل الزمنية-دراسة مقارنة

رسالة تقدم بها

محمد خميس رشيد

إلى

مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل

وهي جزء من متطلبات نيل شهادة الماجستير

في علوم الإحصاء

بإشراف

أ.م.د. أسامة بشير شكر الحنون

المستخلص

إن عملية إختيار متغيرات الانحدار الذاتي في السلاسل الزمنية ذات الابعاد العالية التي تحتوي على عدد كبير من المتغيرات مقابل عدد أقل من المشاهدات تكون عملية صعبة ومعقدة، إذ أن كثرة المتغيرات الذاتية تؤدي الى تعدد الانماط السلوكية للسلسلة الزمنية لذلك يلجأ الباحثون الى التقليل من هذه من المتغيرات وإختيار ذات التأثير الفعلي في النموذج. ان بيانات السلاسل الزمنية في هذه الرسالة تمثل حركة دودة الربداء الرشيقية (Caenorhabditis Elegans (CE وهي عبارة عن مقطع فيديو لحركة دودة طوله ما يقارب (١٤٤) دقيقة وكل دودة تمثل سلسلة زمنية تحتوي على (١٧٩٨٤) مشاهدة بواقع اطارين لكل ثانية. تم إستخدام أسلوب لاسو Least Absolute Shrinkage and Selection Operator (LASSO) وهو أسلوب أساسي يستخدم في إختيار المتغيرات واحد الطرائق الجزائية المستخدمة في إختيار المتغيرات وتقدير النموذج في البيانات العالية الابعاد. كذلك تم إستخدام طرائق اخرى ومنها طريقة LASSO Regularization Generalized Linear Models (LASSO GLM)، اذ يستخدم أسلوب لاسو مع النماذج الخطية المعممة للتخلص من المتغيرات الزائدة والحصول على نموذج جديد نستطيع من خلاله التنبؤ والتحليل. كذلك تم إستخدام أسلوب الانحدار الخطي البيزي لإسلوب لاسو Bayesian linear regression model with LASSO procedure (LASSO BLM) اذ أن نموذج الانحدار البيزي يتعامل مع المعلمات β و σ^2 في نموذج الانحدار الخطي بوصفها متغيرات عشوائية. كما تم إستخدام طريقة الشبكة المرنة Elastic-Net Method إذ إنها طريقة مطورة من أسلوب LASSO مضاف اليه حدي جزاء ℓ_1, ℓ_2 للتخلص من قيود LASSO وبعض ما يعانيه من عيوب والحصول على نتائج أفضل. كذلك تم إستخدام طريقة الشبكة المرنة المعممة Elastic-Net Generalized Linear Models (Elastic-Net GLM) التي تستخدم مع النماذج الخطية المعممة. تم مقارنة طرائق لاسو الثلاثة المختلفة وظهرت النتائج التي تم الحصول عليها هي تفوق طريقة LASSO BLM على طريقتي LASSO و LASSO GLM من خلال قيمة RMSE و معيار AIC وكذلك عدد المتغيرات المختارة. وعند مقارنة نتائج طريقة الشبكة المرنة اظهرت تفوق طريقة Elastic-Net الاعتيادية على طريقة Elastic-Net GLM من خلال الحصول على قيمة معيار RMSE و معيار AIC اقل مقارنة مع طريقة Elastic-Net GLM.

Abstract

The process of selecting the eigen vectors(regression variables) in time series with high dimensions that contains a large , huge number of variables is a difficult process , as the large number of subjective variables leads to multiple behavioral patterns of the time series, so researchers resort to minimize these variables. Choosing the best of them that have an actual impact on the model. The time series data in this thesis was represented by the movement of the *Caenorhabditis elegans* (CE) which is a video clip of the movement of a worm with a length of approximately (144) minutes, and each worm represents a time series containing (17984) observations at two frames per second. The Least Absolute Shrinkage and Selection Operator (LASSO) procedure was used, which is a basic method used in selecting variables and one of the punitive methods used in selecting variables and model estimation in high-dimensional data. Other methods were also used, including the LASSO Regularization Generalized Linear Models (LASSO GLM) method, as the LASSO procedure is used with generalized linear models to get rid of excess variables and obtain a new model through which we can predict and analyze. Bayesian linear regression model with LASSO procedure (LASSO BLM) was also used, as the Bayesian linear regression model deals with the parameters (beta and alpha) and in linear regression model as random variables. The Elastic-Net Method was also used, as a developed method from LASSO, with two penalties added to it to get rid of the limitations of LASSO and some of its defects, and to obtain better results. Also, one of the elastic network methods was used, which is Elastic-Net Generalized Linear Models (Elastic-Net GLM), which are used with generalized linear models. Developed methods such as LASSO GLM and LASSO BLM, as well as the Elastic-Net GLM method, were used to deal with the problem of data multimodality more accurately. The most important results obtained are the superiority of the LASSO BLM method over the LASSO and LASSO GLM methods through the RMSE value as well as the number of

selected variables. Also, the regular Elastic-Net method is superior to the Elastic-Net GLM method by obtaining a lower RMSE standard value compared to the Elastic-Net GLM method.