Ministry of Higher Education and Scientific Research

University of Mosul

College of Computer Sciences and Mathematics

Using the Hybrid Approach of Logistic Regression and Support Vector Machine Methods for Air Pollution Forecasting

Sura Amir Mohammed Nafe'a

M .Sc Thesis

Statistics

Supervised by

Dr. Osama Basheer Shukur

2019 A.D ——————————————— **1441 A.H.**

Abstract

Air pollution studying and forecasting is necessary to control and reduce the damage of environment and human health. Particulate matter (PM_{10}) is a dataset used to measure air pollution. There are many pollutants as sources of air pollution (Co, So_2, O_3, No_x, No, Wind Speed, and Ambient Temperature) may have are effect on PM_{10} variable. PM_{10} and the pollutant variables have been taken from the meteorological station in Kuala Lumpur, Malaysia. All of these variables classified as nonlinear data. Logistic regression (LR) model can be used for modeling and forecasting these multivariable datasets. LR is built using generalized linear model as a special case of linear statistical methods, therefore it may reflect inaccurate results when used with nonlinear datasets. To improve the results of modeling and forecasting, support vector machine (SVM) method has been suggested in this study. LR-SVM hybrid method has been suggested also for more improving of modeling and forecasting accuracy. Time stratified (TS) method in different styles is proposed for satisfying more homogeneity of datasets. TS method includes ordering the similar seasons in different years together as one variable to formulate new variables smoother than their original variables. The results in this thesis reflect outperforming for SVM and LR-SVM methods comparing to LR. LR-SVM has also outperformed both of LR and SVM separately. In conclusion, SVM and LR-SVM hybrid method forecasting can be used for more accuracy with nonlinear multivariate datasets in which PM_{10} is dependent variable.

جامعة الموصل

كلية علوم الحاسوب والرياضيات

استخدام الاسلوب الهجين لطريقتي الة متجه الدعم والانحدار اللوجستي للتكهن ببيانات تلوث الهواء

سرى عامر محمد نافع

رسالة ماجستير

الإحصاء

بإشراف

المدرس

د. أسامة بشير شكر الحنون

م 2019 هـ 1440

المستخلص

تعتبر دراسة بيانات تلوث الهواء والتكهن بها ضرورية للتقليل من اضرارها على البيئة وصحة الانسان والسيطرة عليها. حيث تم استخدم بيانات الجسيمات المعلقة (Particular matter (PM_{10})) لقياس تلوث الهواء. كما ان هنالك عدة ملوثات والتي تعتبر كمصادر لتلوث الهواء مثل (No, No_x, O_3, So_2, Co, وسرعة الرياح و درجة الحرارة المحيطة) والتي قد تؤثر على متغير PM_{10}. ان متغير PM_{10} وبقية المتغيرات الملوثة تم اخذها من احدى المحطات المناخية في كوالالمبور في ماليزيا. تصنف كل هذه المتغيرات غلى اتها بيانات غير خطية كما يمكن استخدام نموذج الانحدار اللوجستي (Logistic regression (LR)) لنمذجة البيانات المتعددة المتغيرات المستقلة والتكهن بها. ان نموذج الانحدار اللوجستي تم انشاؤه بالأساس اعتماداً على النماذج الخطية المعممة و يعد حالة خاصة من الطرق الاحصائية الخطية ولذلك فان نتائج استخدامه مع البيانات الغير الخطية قد تتسم بشيء من عدم الدقة. في هذه الدراسة تم اقتراح طريقة آلة متجه الدعم (support vector machine (SVM)) لتحسين نتائج النمذجة والتكهن . كذلك تم اقتراح الطريقة الهجينة التي تجمع بين نموذج الانحدار اللوجستي والة متجه الدعم (LR-SVM) لتحسين نتائج النمذجة والتكهن يدقة اكبر. كما تم في هذه الدراسة اقتراح طريقة التراصف الزمني (Time stratified (TS)) لتحقيق نتائج اكثر انسجاما وسلاسة مقارنة بالبيانات الاصلية و يتضمن مفهوم هذه الطريقة ترتيب ومراصفة الفصول الموسمية المتشابهة في سنوات مختلفة تباعاً ضمن سلسلة زمنية واحدة. عكست نتائج هذه الدراسة افضلية واضحة لطريقتي SVM و LR-SVM مقارنة بنموذج الانحدار اللوجستي. كذلك تفوقت الطريقة الهجينة على كل من الطريقتين التي تضمنتهما كل على حدا. وتم استنتاج ان تكهنات طريقتي SVM و LR-SVM من الممكن اقتراحهما لتحقيق نتائج اكبر دقة عند استخدامها مع البيانات غير الخطية في حال كان PM_{10} هو المتغير المعتمد.