استخدام خوارزميات التصنيف (Ensemble) في تحليل أداء طريقة الخوارزميات المتعددة

إعداد

عمر هاني محمد الضرغام

إشراف

الدكتور أكرم عثمان المشايخي

الملخص

لقد أصبحت عملية تصنيف البيانات تزداد خلال السنوات الماضية نتيجة الطلب المتزايد على المعلومات من خلال المواضيع والتطبيقات المختلفة.

ولذلك يمكن تعريف عملية تصنيف البيانات بأنها طريقة تصنيف أو ترتيب البيانات بطرق مختلفة مع الأخذ بعين الاعتبار محاور واعتبارات مختلفة بناءً على البيانات المراد تصنيفها.

في هذا الرسالة، سيتم فحص أداء خوارزميات متعددة وتطبيقها على قاعدتين بيانات، الأولى: بيانات الموظفين الذين يغادرون العمل قبل انتهاء ساعات العمل الرسمية، والثانية، حركات الاحتيال على البطاقات الائتمانية الصادرة من املي البطاقات الأوروبيون في عام 2013، وسوف يتم عرض النتائج بناءً على أدوات قياس مختلفة و باستخدام أداة (WEKA) لتصنيف البيانات.

وخلال إجراء عملية التصنيف، تم استخدام مرشح أو منقي (Filter) يسمى (Nominal to Binary Filter) وتطبيق الخوارزميات التالية، (Bagging) و (AdaBoostM) و (Ensemble Selection).

وتمت عملية المقارنة بين هذه الخوارزميات استناداً على أدوات القياس الموجودة في الأداة المستخدمة، ولقد أظهرت النتائج أنه عند تطبيق مرشحات بغير اشراف (Unsupervised Filtering)، تبين أن جميع الخوارزميات المستخدمة لها نفس النتائج عند التطبيق على كل من مجموعات البيانات.

أما عند تطبيق مرشحات مع الأشراف (Supervised Filtering)، فإن النتائج أصبحت مختلفة ولقد تم ترتيبها تصاعديا من الأقل إلى الأعلى كالتالي: (Bagging)، ومن ثم (Ensemble Selection)، وأخيراً (AdaBoostM). أما الترتيب التنازلي فهو، (AdaBoostM)، ومن ثم (Ensemble Selection)، واخيرا (Bagging).

وفي النهاية، نستعرض الملخص للبحث والنتائج النهائية التي تبين من الأفضل في هذه الخوارزميات بناءً على الأدوات القياسية المستخدمة.

Using Ensemble Classification Algorithms in Performance Analysis of a Multi Algorithm Methods

Prepared by:

Omar Hani Mohammad Al-Dorgham

Supervised by:

Dr. Akram Othman Al- Mashaykhi

Abstract

Data classification has been growing through years due to the highly demand of the information needed for different subject and in different applications.

Data classification is a way or a procedure to classify data into different types with corresponding to different perspectives according to the need of data. In this thesis, a research model is proposed for the classification in terms of measure the best performance algorithm applied to same databases which depends on different measurement tools in terms of implementing the multi method algorithms technique, where the process will be done by using WEKA tool. In the processing of this model, two data bases are used, Human Resources Analytics, and Credit Card Frauds by applying firstly,

Nominal to Binary filter and secondly, the used algorithms, which are Bagging, AdaBoostM, and Ensemble Selection algorithms. Finally, the comparison step, where the results of the previous steps will be compared in terms of measurement tool, which are Kappa Statistics, Mean Absolute Error & Root Mean Square Error, Relative Absolute Error and Root Relative Squared Error.The Experiment showed that when applying the unsupervised filtering method for all algorithms, the results will be the same for both databases. On the other hand, when applying supervised filtering method, the results are different, where the order of the algorithms in an ascending order, starting from the lowest to highest, is AdaBoostM, then Ensemble Selection, and finally, bagging algorithm. Furthermore, when using descending order, the results will be, firstly, Bagging, then Ensemble Selection, and finally, AdaBoostM.