University of Science and Technology

College of Post Graduate Studies and Academic Advancement

Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of science in Information Systems

Comparing between machine learning tools (rapid miner and weka)

By:

Raghda Mohamed Elhassen Mustafa

Supervisor:

Dr. Atif Ali

2014

Abstract

Many data mining and machine learning tools and software are available for everyone and different usage such as the Waikato Environment for Knowledge Analysis (WEKA), Rapid Miner, These tools and software provide a set of methods and algorithms that help in better utilization of data and information.

This research explains a comparative study between two of the free available data mining and machine learning tools, and used j48 algorithm to make this comparative study and measure accuracy, recall and precision after implement j48 algorithm in weka and rapid miner tool.

The main objective of this research is Apply j48 algorithm in two tools (weka and rapid miner) to Measure accuracy, recall and precision for this algorithm and compare the result of our experiment to discuss the differences between performance of j48 algorithm in weka and rapid miner.

The followed methodology here is to study two of free data mining and machine- learning tools to be tested by selecting j48 classification algorithm to test these tools' performance. And measure their performance through accuracy, precision and recall and compare the results.

After execution J48 algorithm in the two tools (Weka, Rapid Miner) in term of Accuracy, Precision and Recall using the first dataset we found that the result was different, But in the second and the third dataset the result was the same in the two tools.

At last I notice that when the dataset is big the rapid miner gave me higher result in all terms.

المستخلص

في الوقت الحالي ومع كثرة البيانات والأدوات الخاصه باستخراج وتعدين هذه البيانات والتي صارت متاحة للجميع صارت هناك مشكله في معرفه الاداه الفعاله التي توفر مجموعة من الطرق والخوارزميات التي تساعد في تحسين الاستفادة من البيانات و المعلومات.

هذا البحث يوضح مقارنة بين اثنين من أدوات استخراج البيانات و التعلم الآلي المتاحه مجانا وهاتين الاداتين هما (rapidminer و weka) حيث تم استخدام خوارزمية j48 في صنع هذه الدراسة والمقارنة بين الاداتين عن طريق قياس الدقة و الاستدعاء.

الهدف الرئيسي من هذا البحث هو تطبيق خوارزمية j48 في أداتين (WEKA و rapidminer) ل قياس الدقة والاستدعاء لهذه الخوارزمية في الاداتين.

و المنهجية المتبعة هنا هي دراسة اثنين من ادوات استخراج البيانات المجانيه ، لفحصهما عن طريق اختيار خوارزمية تصنيف j48 لاختبار أداء هذه الأدوات " . وقياس أدائها من خلال الدقة و الدقة والمراجعة و مقارنة النتائج .

بعد تطبيق خوارزمية J48 في الاداتين باستخدام مجموعة البيانات الأولى وجدنا أن النتيجة كانت مختلفة، ولكن في مجموعه البيانات الثانية و مجموعة البيانات الثالث كانت النتيجة هي نفسها في الأداتين.

أخيرا لقد لاحظت أنه عندما تكون مجموعة البيانات كبيره الاداه rapidminer اعطتنا نتائج اعلي في جميع المقاييس.

1.1 Introduction

Every day the systems grown up more and more and its data get bigger and bigger so day by day we get a huge data in our institutions and organizations (Oracle DB, SQLDB …), all companies and organizations are going towards the system Integration, but still managers on the top level cannot get benefits from these huge data integrations.

Machine Learning widely used in computer science and other fields .It used to make better decision making by detect pattern in data and it looking at the data to extract useful information for use in predicting the future.

Many data mining and machine learning tools and software are available for everyone and different usage such as the Waikato Environment for Knowledge Analysis (WEKA), Rapid Miner, These tools and software provide a set of methods and algorithms that help in better utilization of data and information.

We can use some of data mining technique to help us to make better decision making.

1.2 Problem Definition

The main problem in this research determine the best tool between WEKA and Rapid Miner by providing a comparative study between two tools which done by applying J48 algorithm on each tool, and select the best tool depending on the final results.

1.3 Main Objective

The main objective of this research is:

Apply j48 algorithm in two tools (weka and rapid miner) to Measure accuracy, recall and precision for this algorithm.

1.4 Sub objectives

- To apply j48 decision tree algorithm.
- To compare the result of our experiment to discuss the differences between performance of j48 algorithm in weka and rapid miner.

1.5   Methodology

The methodology of the study content two of free data mining and machine-learning tools to be tested by selecting j48 classification algorithm to test these tools' performance. And measure this performance through accuracy, precision and recall and compare the results.
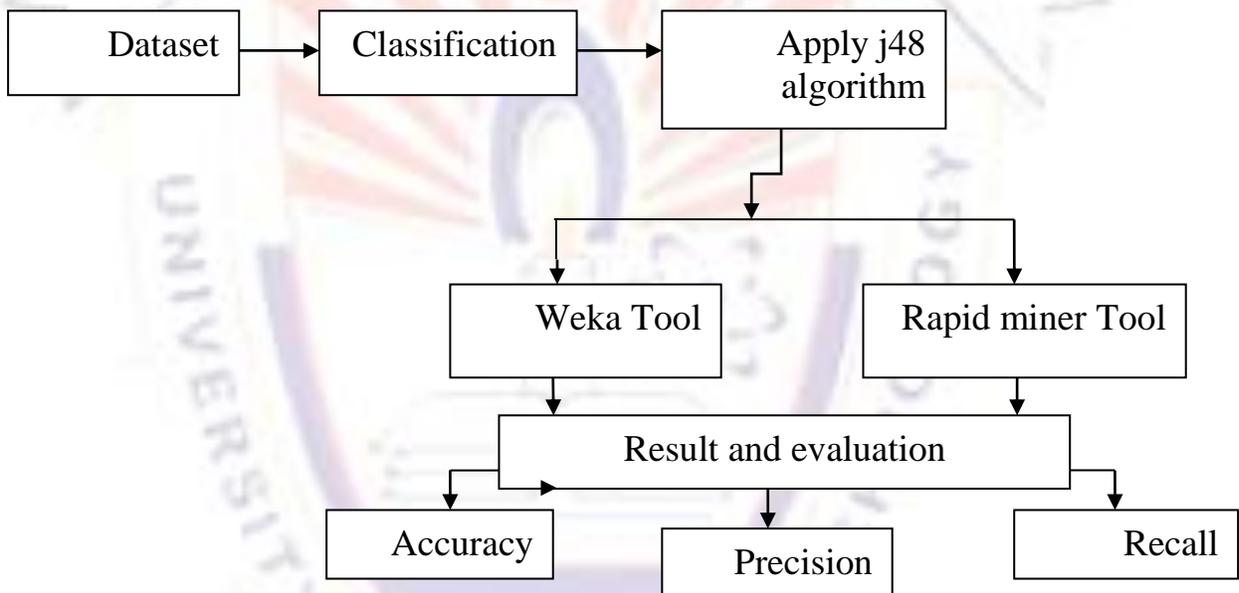

Figure 1.1 Methodology of research

1.6 Research Structure

- Chapter two: Background for machine learning.
- Chapter three: Classification and decision tree.
- Chapter four: Experimental and results.
- Chapter five: Conclusions and recommendations