بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Graduated Studies

Classification Algorithms Comparison- Case Study: Cancer Patients(SEER Data Set)

مقارنة خوارزميات التصنيف - دراسة حالة: مرضي السرطان

By

Musab Abd algader Abd alkreem

Supervisor

Dr. Mohamed Elhafiz Mustafa Musa

2014

ACKNOWLEDGEMENT

Praise is to Allah, I would like to express my gratitude to many people who helped me in different ways with the development of this research. Without their continuous support and guidance, the completion of my research would be impossible. I wish to express my most sincere thanks to my supervisor Dr. Mohamed Elhafiz Mustafa Musa, who had a major role in ending this research in this way and gave me a lot of help and assistance, it is my luck to study and work under his guidance. Many thank to my family for the many years' caring and support. Finally, I would specially like to thank friends for everything they have done for me.

ABSTRACT

Data mining is the automatic search of huge data to discover patterns and trends that go beyond simple analysis. Data mining is also known as Knowledge Discovery in Data (KDD). This study investigates the discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. To do such investigate a large data set needed one of these data sources is SEER[1] (Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. In this study we have investigated three data mining techniques Multilayer Perceptron (MLP), K-nearest neighbor and the C4.5 decision trees the goal is to find the best accuracy to predict 5 years survivability of breast cancer.

SEER database (period of 1973-2009 with 657,712 records) were used, starting from previous study we determined common variables use, after preprocessed there are 18 variables and 180,302 records.

Weka was used to train and test the three techniques. The result show that the best technique is C4.5 accuracy is %95.6 and the second technique is K-nearest neighbor with accuracy %95.4 and the worst is MLP with accuracy %95.3.

المستخلص

تنقيب البيانات هو البحث التلقائي في بيانات ضخمة لاكتشاف الأنماط والاتجاهات. وكما هو معروف تنقيب البيانات جزء من اكتشاف المعرفة في البيانات (KDD). تبحث هذه الدراسة في المقارنة بين ثلاثة مصنفات (Classifiers) اكتشاف معدل البقاء على قيد الحياة عن طريق استخراج المعرفة من البيانات المتعلقة بهذا المرض. للقيام بمثل هذا التحقيق مجموعة كبيرة من البيانات المطلوبة واحدة من هذه المصادر البيانات SEER [1] (مراقبة الأوبئة والنتائج النهائية)، والذي هو مورد فريد وموثوق بها والضرورية للتحقيق في جوانب مختلفة من السرطان. في هذه الدراسة قمنا بالتحقيق ثلاث تقنيات استخراج البيانات متعدد الطبقات المتعرف ((MLP، K-أقرب جار والأشجار قرار C4.5 الهدف هو العثور على أفضل دقة للتنبؤ ٥ سنوات البقاء على قيد الحياة لسرطان الثدي. واستخدمت قاعدة بيانات برنامج سير (فترة ١٩٧٣-٢٠٠٩ مع ٦٥٧٧١٢ سجلات)، بدءا من دراسة سابقة قررنا المتغيرات المشتركة استخدام، بعد preprocessed هناك ١٨ المتغيرات و١٨٠٣٠٢ السجلات. كان يستخدم لتدريب ويكا واختبار التقنيات الثلاث. تظهر النتيجة أن أفضل أسلوب هو دقة C4.5 هو ٩٥.٦٪ والأسلوب الثاني هو K-أقرب جار مع دقة ٩٥.٤٪ والأسوأ هو MLP بدقة ٩٥.٣٪.

Table of Contents

Chapter 1

Chapter 2

Chapter 3

LIST OF FIGURES

LIST OF TABLES