UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Comparative Study of Classification Algorithms

(J48 and Simple CART)

by

Yahia Abd Elhamid Ahmed Elhag

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements in Partial Fulfillment of the Requirement for the Degree of Master of Computer Science

Supervisor

Dr. Atif Ali Mohamed

April 2016

Abstract

There is a huge amount of information locked up in database information that is potentially important but has not yet been discovered or articulated. The problem then becomes how to analyze the data. Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

Classification is the process of finding a model or functions that describe or distinguish concepts or classes of data that aims to predict the class of an unknown object label. Classification comes with variety of techniques like decision tree, neural network, genetic algorithm and K-nearest neighbour. The decision tree is one of the most popular techniques used for classification and applicable to data whose values are known precisely.

In this research, data mining classification algorithms are used to categorize the thyroid disease as sick or negative and our aim is to find out which of the data mining models is the most suitable model that can be used for the thyroid disease classification according to the model results.

We conducted a comparative study between J48 and CART classification algorithms. The experiments were conducted in WEKA involves both algorithms and testing of set of data related to the medical field with 10-fold cross validation. The data set in the experiment was divided into three different sizes of records first with 1772 then 2772 and for the last with the all 3772 records. From the investigation and comparison result, it can be said that J48 classification method is better than CART in small to medium size data sets and it has a low computation cost and faster training time, but with big and huge data sets CART is better but takes more time to build model.

المستخلص

تخزن الحواسيب كمية هائلة من البيانات داخل قواعد البيانات، مع احتمال أن تحتوي هذه البيانات على معلومات مفيدة ومهمة لم يتم اكتشافها، اصبح من الضروري إيجاد وسيلة لتحليل هذه البيانات وإستخراج معلومات مفيدة منها مما أدى إلي ظهور علم التنقيب عن البيانات وأدواته المختلفة، حيث أن تنقيب البيانات يعني إستخراج المعلومات الضمنية التي لم تكن معروفة سابقاً والتي يمكن أن تكون مفيدة من البيانات.

التصنيف هو عملية إيجاد نموذج أو ألية تصف أو تميز المفاهيم أو الفئات من البيانات التي تهدف إلى التنبؤ بفئة كائن من تصنيف غير معروف. التصنيف يأتي مع مجموعة متنوعة من التقنيات مثل شجرة القرارات (decision tree)، الشبكة العصبية (neural network) ، الخوارزمية الجينية (genetic algorithm) وتقنية الجار الاقرب (K-nearest neighbor). شجرة القرارات هي واحدة من التقنيات الأكثر شعبية المستخدمة للتصنيف وتطبق على البيانات التي تعرف قيمها على وجه التحديد.

في هذا البحث، تم استخدام خوارزميات التنقيب عن البيانات لتصنيف أمراض الغدة الدرقية اما مريض او نتيجة سلبية وهدفنا هو معرفة أي من نماذج التنقيب واستخراج البيانات هو النموذج الأنسب التي يمكن استخدامها لتصنيف أمراض الغدة الدرقية وفقا لنموذج النتائج.

أجرينا دراسة مقارنة بين خوارزميات تصنيف وهي J48 وCART. وأجريت التجارب في WEKAأنها تنطوي على كل من الخوارزميات وتم الاختبار على مجموعة من البيانات المتعلقة المجال الطبي مع استخدام تقنية 10-fold cross validation. تم تقسيم مجموعة البيانات في التجربة إلى ثلاثة احجام مختلفة من السجلات أولاً مع 1772 سجل ثم 2772 و اخيراً مع كل السجلات وهي 3772 سجل . من التحقيق ونتيجة المقارنة، يمكن القول أن طريقة تصنيف J48 أفضل من CART في مجموعة البيانات صغيرة ومتوسطة الحجم ولها وقت تدريب اسرع و تكلفة حساب منخفضة، لكن خوارزمية CART هي الافضل مع مجموعة البيانات الكبيرة مع انها تستغرق وقت اطول في التدريب و بناء النموذج.

1.1 Introduction:

Nowadays, we cannot imagine our life without computer programs and the fact is that they have become so important that nothing can replace them. Computer programs would not exist without algorithms and with computer applications becoming indispensable in almost all aspects of our professional and personal lives, studying algorithms becomes a necessity for more and more people. After all, algorithms can be seen as special kinds of solutions to problems not just answers but precisely defined procedures for getting answers.

Comparative studies of algorithms are the act of comparing two or more algorithm with a view to discover something about one or all of the algorithms being compared. This technique often utilizes multiple disciplines in one study. The analysis of algorithm defines that the estimation of resources required for an algorithm to solve a given problem. Sometimes the resources include memory, time and communication bandwidth.

In conducting comparative studies, classification researchers and other data miners must be careful not to rely too heavily on stored repositories of data as its source of problems. The comparative study, does not usually propose an entirely new method;most often it proposes changes to one or more known algorithms, and uses comparisons to show where and how the changes will improve performance [17].

The important of algorithm comparing is to find a suitable algorithm for a specific problem. One way to recognize the best suitable algorithm for a given problem is to implement both algorithms and find out their efficiency, most importantly the running time of a program [9].

Comparison in this research is made on accuracy, sensitivity and specificity using true positive and false positive in confusion matrix generated by the respective algorithms. Also we can use the correct and incorrect instances that give us a most efficient method for classification by using the confusion matrix.

1.2 Problem definition:

Since we are in the age of data and data mining becomes a major science in our life we need to find what the best to do that, and in this research we are seeking to find the best classification algorithm for numerical data sets.

1.3 Main objective:

Is to make a comparison between classification algorithms to come out with a suitable model that can be used for classify numerical data sets providing the best result accuracy and takes less time to be built.

1.4 Methodology:

This section briefly explains the methodology adopted in this research, it is also discusses the research steps taken in comparing J48 and CART which are:

1. Choose the algorithms to include in the comparison (J48 and CART).
2. Choose the suitable data set for testing. For example, if the algorithm is supposed to handle large attribute spaces, choose a data set with a large number of attributes.
3. Divide the data set into 10 subsets for cross validation other values may be chosen depending on the data set size.
4. Run a cross validation.
5. Run test multiple times with different number of instances.
6. Summarizes the test results and discuss it.

1.5 Research structure:

The research structures consist of the five chapter are the following:

In chapter two background of the data mining and the data mining techniques and applications, In chapter three the classification techniques and methods and, In chapter four implementation of simple cart algorithm and J48 using WEKA tools for data mining and what founded or results on this implementation, In chapter five the Conclusions and recommendations.