UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Apply Clustering Algorithms In Optometry Using Attribute Selection Method

By:

OmimaAbdeElrahmanMarouf

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements

in PartialFulfillment of the Requirement for the Degree of Master of Information Technology

Supervisor

Dr. AbdelhamidSalih

January 2017

Abstract

Data mining is the practice of examining large pre- existing databases in order to generate new information. It is the process of selecting, exploring, and modeling large amounts of data to discover unknown patterns or relationships useful to the data analysis.

Data Mining is also useful and important in the medical field which can be used on the determination and prevention of diseases and health care related issues.

There are several different kinds of data mining techniques are available such as classification, clustering, association rules and select attribute.In this research we used the method of "clustering" which is used to detect the grouping in many status. Clustering has many algorithms and here we used only two: The simple K-means: is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean , and the DBSCAN: Density-Based Spatial Clustering of Applications with Noise algorithm are capable of discovering clusters of arbitrary shapes. This provides a natural protection against outliers. And to apply this two algorithms we used the applications "Weka" to evaluate the accuracy timing and general performance of both of this algorithms. And to applied on the The OPTOMETRY dataset. The result after the comparisons process was that k-means algorithm is best algorithm compared with other clustering algorithm (dbscan) so it will be most suitable for the users.

المستخلص

استخراج البيانات (data mining) هو ممارسة فحص قواعد البيانات من أجل توليد معلومات جديدة . وهو عملية اختيار، واستكشاف، ونمذجة كميات كبيرة من البيانات لاكتشاف أنماط غير معروفة أو علاقات مفيدة لتحليل البيانات.

تنقيب البيانات (data mining) مهم في المجالات الطبية ويتم استخدامه في عملية التحليل والتنبؤ والاكتشاف والوقاية.

هناك العديد من تقنيات التنقيب classification, clustering, association rules and select attribute. وفى بحثنا هذا استخدمنا تقنية التجميع (clustering) ومهمته اكتشاف التجمعات في مجموعه من الحالات وتقييم الاداء بين خوارزميات التجميع المختلفة والتجميع لديه كثير من الخوارزميات وفي هذا البحث استخدمنا خوارزميتين:

Simple k- means algorithm and DBSCAN (Density-Based Spatial Clustering Applications with Noise algorithm).

وتم تطبيق الخوارزميتين باستخدام برنامج الويكا علي مجموعه من البيانات الحقيقيه في مجال البصريات وتقييم الاداء لكل من الخوارزميتين من حيث الوقت والدقة والاداء العام.

وكانت النتيجة التي تحصلنا عليها من تطبيق الخوارزميتين ان الخوارزمية k_means هي افضل من حيث الدقة والاداء والوقت مما يجعل استخدامها أفضل للمستخدم .

1.1 Introduction

Data mining technologies were developed in order to extract meaningful information from these large stores of data taking a place in the overall field of knowledge Discovery methods where by new information is derived from a combination of previousknowledge.

Modern medicine generates a great deal of information stored in the medical database, extracting useful knowledge and providing scientific decision- making for the diagnosis and treatment of disease from databaseincreasingly become necessary.

Data mining in medicine can deal with this problem and also improve information in the hospital and development community medicine.[14]

Information Technology health care provide significant potential to improve productivity and quality of patient care. The area of Data mining in health care is growing rapidly because of strong need for analyzing the vast amount of clinical data bases stored in Hospitals.[14]

This searchdescribes applications of data mining for the analysis of optometry data.The data captured by health clinics from many patient records in the database contained errors and missing values. In addition, many of the records were not in a form suitable for data mining; they had to be transformed to more meaningful attributes. Here the challenge of datamining techniques appears. A major problem in medical science is in attaining diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification of large-scale data. Data mining has played an important role in optometry because it can unearthhidden knowledge from huge amount of eyesdiseaserelated data.

1.2 Problem Statement

In this research we investigateclustering data mining algorithms, and answer the question: what is the best clustering algorithm that produces best results on aspect of accuracy and running time? We have used data measurements revealed Consideration optics to check into our search.

1.3 Research objectives:

The goal with this research can be is divided as following:

1. To build dataset of optometry.
2. To investigate the (simple K-means and DBSCAN clustering) algorithm to build clustering model in optometry.
3. To evaluate the results of the investigation.
4. To comprise between the algorithms

1.4 Research methodology:

The dataset used in this study were collected from Mecca eyes hospital in Omdurman of patient's records about 1530 cases females and males of all ages and classes Sudanese. We are using clustering technique using weka .We will the testing of algorithms first without attribute selection, and then after that will be testing with attribute selection. And compare between their performance and accuracy of each one at the time of the implementation period.

1.5 Research Organization:

This search contains five chapters: chapter two the literature review, It is describing a set of concepts on the subject of search and discussed data mining background , and contain a related work.Chapter three contains the data collection techniques and data mining techniques and summarize simplified algorithms used in this research and the tools used.Chapter four to conduct experiments on two algorithm (simple K-means and DBSCANclustering) in tow case without attribute selection and in another case with attribute selection, and analysis of the results, then evaluate the results, finally the comparison between the results of all algorithms.Chapter five is the last chapter, defines conclusions and recommendations.