University of Science and Technology

Faculty of Post Graduate Studies & Academic Development

Applying K-mean Algorithm to Cluster Student Assessment Data

Thesis submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Information System

By : Aisha Abdelgader Mohamed

Supervisor : Dr. Atif Ali Mohamed

February - 2014

Abstract

Improving student's academic performance is not an easy task for the academic community in higher education. The academic performance of computer science students in their first university year at is a turning point in their educational path.

In this research apply one of the approaches of data mining which is clustering it's data mining techniques that is the process of partitioning a given data set into groups of similar objects, clustering in this research applied on the student assessment data for Senate at UST batch 2008-2009 by used one of data mining tools named WEKA that is open source software, it's easily usable by people who are not data mining specialists, though K-mean algorithm one of the famous clustering algorithms, which is well known their efficiency in assembling large data sets; the goal is to divide the students into k clusters such that students lying in one cluster should be as close as possible to each other's (homogeneity) and students lying in different clusters are further apart from each other

The results that were obtained After applied K-means clustering algorithm, the students can be grouped in two classes high(GPA between 100-60) that include four estimates (90-100 excellent, 80-89.9 v-good, 70-79.9 good, 60-69.9 pass); or low(GPA between 59.9-0) that include two estimates of student (50-59.9 low, 0-49.9 failure) based on their new grade.

المستخلص

تحسين الأداء الأكاديمي للطالب ليست مهمة سهلة بالنسبة للمجتمع الأكاديمي في التعليم العالي. الأداء الأكاديمي لطلاب علوم الحاسوب خلال السنة الأولى في جامعتهم هو نقطة تحول في المسار التعليمي.

في هذا البحث تم تطبيق احد مفاهيم تنقيب البيانات هو التجميع وهو تقنية لتقسيم مجموعة معينة من البيانات إلى مجموعات من الكائنات مماثلة.التجميع في هذا البحث طبق على بيانات تقييم الطلاب لمجلس الأساتذة في جامعة العلوم والتقانة، قسم علوم الحاسوب الدفعة 2008-2009 باستخدام احدى أدوات إستخراج البيانات تسمى WEKA وهي من البرمجيات مفتوحة المصدر وسهلة الإستخدام بواسطة أي شخص ليس متخصص في مجال تنقيب البيانات، عبر الخوارزمية K-mean وهي من خوارزميات التجميع المشهورة ومعروفة جيداً بكفاءتها في تجميع مجموعات كبيرة من البيانات والهدف هو تقسيم الطلاب إلى مجموعات ; الطلاب في مجموعة واحدة يجب أن تكون صفاتهم أقرب ما يمكن إلى بعضها البعض (متجانسة) والطلاب في مجموعات مختلفة صفاتهم متباعدة عن بعضها.

النتائج التي توصلنا اليها بعد تطبيق خوارزمية التجميع k-mean هي تقسيم مستوى الطالب حسب درجاته إلى مستويين أحدهما مستوى عالي ويتضمن أربعة تقديرات وهي في حالة حصوله على 90-100درجة فتقديره ممتاز، 80-89.9 درجة فتقديره جيد جدا، 70-79.9 درجة فتقديره جيد، 60-69.9 درجة فتقديره مقبول. والآخر مستوى متدني ويتضمن تقديرين في حالة إحرازه على 50-59.9 درجة فتقديره ضعيف ،0-49.9 درجة فتقديره رسوب.

1.1 Introduction

There is strong perception of the need for analysis Big Data as well the benefits it can bring by it and the methods to achieve success. which will have measurable long-term impact on institutions.

Graded Point Average (GPA) is a commonly used indicator of academic performance. GPA still remains the most common measure used by the academic planners to evaluate progression in an academic environment. many measures could act as barriers to student attaining and maintaining a high GPA that reflects their overall academic performance, during their tenure in University. These measures could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance.

This research describes and explains clustering in Big Data, data clustering application, and finally implements a k-means algorithm on open source WEKA software to cluster student data that helps Senate in University of Science and Technology, to improve the performance of students and assess the level of students before the final exams.

1.2 Research Problems

- The Student academic data stored at random on record keeping and increasing rapidly.
- Senate cannot assessment student before final exams for semester, because it is not divided this data into levels according to student's degrees.
- Examination Center does not benefit from this data in decision making only after divided according to student levels.
- These data written in Arabic language but WEKA software do not take the Arabic language and also there is spaces between each word, but the WEKA

software do not deal with these voids only after work breaks between each word and the other.

1.3 Research Objectives

The objectives of this research are:

- Study and discuss machine learning methods in big data, and to review the previous studies in this field.
- Implement one of the famous clustering algorithms using student's assessment data at UST Department of Computer Science by using WEKA software.
- Discuss clustering, clustering algorithm of Big Data
- This information can help Senate to improve the performance of students and assess the level of students.
- Examination Committee and Senate identified their student’s behaviors and levels to retain valued students before the final exams.

1.4 Research Methodology

More research methods importance for a human being is experimental method because this approach helped build his or her own development and through observation and experimentation, and access to right results and see peaceful ways to deal with the phenomena and interpretation.

Best Scientific Research methods because this approach depends mainly on scientific experiment, which provides an opportunity to learn the process of finding and enact laws through these experiences.

This approach has become far incomplete images and is used in a manner that depends on the basis scientific rules. The apparent value of the experimental method in Pure & Applied Sciences.

1.5 Research Organization

This research organized as follows:

Chapter two define Big Data, Big Data importance, applications, technology is also discussed in this Chapter which describes different types of Big Data technology and principles, Big Data problems, challenges and future. Chapter three for different results, where we research and Machine learning methods for clustering of large data cluster of massive data cluster analysis, cluster models, clustering and association rules, clustering algorithms and the applications of cluster algorithms. Chapter fourth explains of environment of WEKA tool, follows by using K-means algorithm of cluster of Big Data implemented on WEKA software. Finally Chapter fifth conclusion of this research and the recommendations to be taken into account for the following studies.