University of Science and Technology

Faculty of Computer Science and Information Technology

Thesis submitted in fulfillment of the

requirements for the degree
of Master of Computer Science

A COMPARATIVE STUDY OF ECM, K-MEANS AND FARTHEST FIRST CLUSTERING ALGORITHMS

by:

MAHALA ELZAIN BERAIMA

Supervisor :

Dr / Atayeb Altaher

May 2013

Abstract

Clustering is one of the most important techniques of data mining. Clustering technique in data mining is an unsupervised machine learning algorithm that finds the groups of object such that objects in one group will be similar to one another and are dissimilar to the objects belonging to other clusters. Clustering is called unsupervised machine learning algorithm as groups are not predefined but defined by the data. So the most similar data are grouped into the clusters. In this thesis, we compared the performance of three clustering algorithms namely ECM, K-means and Farthest First for the purpose of clustering and we used the implemented K-means and Farthest First algorithms in Weka software and we used the implemented ECM algorithm in Matlab software. To compare between the three, we used standard KDD 99 benchmark data sets. The results Farthest First algorithm are compared with the results of k-means clustering according to classification performance and The results ECM algorithm are compared with the results of k-means clustering according to objective function value. It shows that ECM clustering algorithm outperform the K-means and Farthest First algorithms and achieved better clustering accuracy as per the results of clusters evaluation, hence the consistent superiority of ECM over both Farthest First and K-means.

الخلاصــة

التجميع عبارة عن بيانات وصفية و هي واحدة من أهم طرق تعدين البيانات و طريقة التجميع تعتبر من خوارزميات آلة التعلم التى لا تحتاج الى تدريب ، و هي عبارة عن عملية تنظيم أو تصنيف أو تقسيم الأشياء الى مجموعات بناءً على أوجه الشبه ، الأشياء الموجودة فى مجموعة واحدة متشابهه مع بعضها البعض و مختلفة عن أي مجموعة أخرى . التجميع يسمى خوارزميات آلة التعلم التى لا تحتاج الى تدريب ويكون تصنيف الأشياء فيها ليست مسبق بل يحدد بناءً على أوجه الشبه من قبل العينة أي تركيبة حقول البيانات المختارة المراد إجراء الدراسة عليها. و فى هذا البحث لقد قمنا بإجراء دراسة مقارنة أداء ثلاثة من خوارزميات و هي بالتحديد ئي سي ام ، كي مينس و فارذست فيرست بغرض التجميع و معرفة أيهما أفضل و تفوقه عن غيره و قمنا بتطبيق كي مينس و فارذست فيرست فى الويكا و ئي سي ام فى الماتلاب . و للمقارنة بين تلك الخوارزميات لقد تم أختبارها جميعاً بواسطة نفس البيانات القياسية المكتشفة للمعرفة و مقارنةً لنتائج كي مينس و فارذست فيرست وفق تصنيف الأداء من حيث الدقة و السهولة و المرونة و نتائج ئي سي ام و كي مينس وفق قيمة دالة الهدف(أقل قيمة أفضل تجميع) و بناءً على نتائج تلك التجارب التى أجريت و لقد أتضح جلياً بأن خوارزمية ئي سي ام هي الأفضل من حيث الدقة و ثبات تفوقها على غيرها .

AN INTRODUCTION

1 Background

Clustering is a descriptive data mining task that aims to identify homogeneous groups of objects, based on the values of their attributes. It is particularly useful in problems where there is little prior information available about the data, and a minimum number of assumptions are required. Clustering is appropriate for the exploration of interrelationships among the data points when assessing their structure.

Many organizations depend on Data Clustering for making decisions and judgment on large volumes of dynamic network data. Networks can be monitored based on methods such as statistical intrusion and abnormal detection for attacks or malicious. The old saying that “a picture is worth a thousand words” often understates the case, especially with regard to moving images, as our eyes are highly affected by evolution to interpreting a movement and clustering the changes of surrounding. Therefore, network monitoring is an important demanding task. The task is even more complicated when dealing and working with highly dynamic information. However, reduction of the complicated network traffic data into simple information (representive sample) and visualize it into a suitable platform are significant challenges for a network administrator.

Our research works have been devoted to contribute to the subject area of data clustering. Data clustering acts as intelligent tools, a method that allows the user to handle large volume of data effectively. The basic function of clustering is to transform data of any origin into a more compact form, one that represents accurately the origin data. The compact representation should allow the user to deal with and utilize more effectively the original volume of data. Clustering techniques are normally utilized to determine a possible attack, due to the uncertainty nature of

intrusion. The accuracy of the clustering is vital because it would be counter-productive if the compact form of the data does not accurately represent the original data. Many research works have technique in clustering inside the networks. Thus far, most of the work was done in offline mode with requires data collection, analyzing and profile creation phase to be completed first. The drawback of this offline approach is its static nature. Static means if changes happened to the network behaviors, all the phase need to be repeated again in order to adapt with new network characteristics.

One of our main contributions is addressing the investigation performance properties of our algorithms ECM, K-means and Farthest First Algorithm in order to compare in terms of performance (i.e. convergence speed) and accuracy (i.e. quality clusters formation).

1.1 Machine Learning and Data Clustering: An Overview

Data clustering is used to grouping unlabeled patterns into clusters based on similarities. There are three general basic steps for data clustering pattern representation, similarity measure to use and clustering or grouping. Clustering also used to rectify an anomaly. Any action that significantly deviates from the normal behavior , referred to as outlier detection refers to detecting patterns in a given data set that do not conform to an established normal behavior. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. Anomalies are also referred to as outliers, change, deviation, surprise, aberrant, peculiarity, intrusion, etc. Anomaly detection based on the normal behavior of a subject. Sometime assume the training audit data does not include intrusion data. Any action that significantly deviates from the normal behavior is considered the intrusion.

In particular, machine learning is an area of information science concern with the creation of information models from data, with the representation of knowledge, and with the elucidation of information and knowledge from processes and objects. Machine learning includes methods for feature selection, model creation and knowledge extraction. There are three broad categories of machine learning techniques exist. Unsupervised method also called data clustering, grouping unlabeled patterns into clusters based on similarities detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised method is to build a predictive model(classifier) to classify or label incoming patterns, require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

1.2 Problem Statement

Clustering is one of the most important techniques of data mining and there are various clustering algorithms for different purposes.

The research problem is concerned with, how to find the better clustering algorithm among comparison of three clustering algorithms namely ECM, K-means and Farthest First.

1.3 Research Objectives

The main objectives of this research work are to study and investigate the evolving clustering (ECM), K-means (KM) and Farthest First algorithms and to identify their strengths

and weaknesses. The investigation includes using The KDD Cup 99 Intrusion Detection Evaluation data as the benchmark to test.

To evaluate the performance of three clustering algorithms ECM, K-means, and Farthest First is also the purpose of this research work, through some comparisons in assessing the quality of a clustering method in terms of the objectiveness, efficiency and validity to show the reliability of the algorithms in clustering data.

1.4 Contribution of This Thesis

The main contributions of this research are towards the following:

- We studied and investigated the ECM/KM and FFA algorithms thoroughly and identified its main strengths and weaknesses in clustering data.
- We also investigated performance properties of our three algorithms through summaries of comparative analysis in order to compare in terms of performance.

1.5 Thesis Organization

This thesis is organized into six chapters .The contents are arranged such that each previous chapter provides a basic idea to further proceed to the next chapter. Firstly in this chapter the background principles of data clustering and an overview of machine learning were covered along with problem statement, research objectives and contributions.

In Chapter 2, discussed literature and fundamental concept related to research work and issues surrounding it. In general, algorithms ECM, K-means and Farther First were defined and described. Some basic uncommon definitions used throughout this thesis are also introduced.

Chapter 3 is the main chapter in this thesis. The WEKA and NeuCom model engine is introduced in this chapter, which covers the methodology of how the research is carried out.

Chapter 4 consists of implementation details and issues of algorithms. Also illustrate the experimentation had done with The KDD Cup 99 Intrusion Detection Evaluation data as the benchmark data. Both experiment details are described.

In-depth analysis and discussion of the results from both experiments described in Chapter 4 are the primary content of Chapter 5. This chapter is divided into two parts. The first part reports the results from the each experiment using benchmark data, and the discussion is presented at the end of both experiment results.

Chapter 6 summarizes this thesis. Revisit research contribution with regard to methods that proposed in chapter 3 and its result in chapter 5. Finally, a discussion and suggestion for future work directions pertaining to this research is presented.