Abstract—The vast amount of information hidden in huge databases has created tremendous interests in the field of data mining. This paper examines the possibility of using data clustering techniques in oral medicine to identify functional relationships between different attributes and classification of similar patient examinations. Commonly used data clustering algorithms have been reviewed and as a result several interesting results have been gathered.

Keywords—Oral Medicine, Cluto, Data Clustering, Data Mining.

I. INTRODUCTION

In this paper, potential pitfalls and practical issues about data mining in oral medicine using data clustering techniques are discussed. In oral medicine, theoretical education to dental students is usually given through lectures, books and scientific papers. Text books often present a small number of cases for each diagnosis. The information students receive may therefore not reflect the reality a clinician in oral medicine encounters in daily practice. When the students graduate the learning that comes with experience from treatment outcomes may therefore be missing. mEduWeb is a program that was written and designed earlier to give students the possibility to study oral medicine through a web interface. mEduWebII used Medview database which contains several thousand patient examinations [1]. The purpose of our work has been to seek improvements in the current mEduWebII program. The objective is to explore data clustering techniques for finding patient examinations that are similar to each other. Several interesting and useful results have been gathered through a series of experiments.

Clustering is the unsupervised classification of patterns into clusters [2]. Finding interesting patterns in large datasets has attracted considerable interest recently [5]. Clustering classifies similar objects into different groups [3]. An example of clustering is shown in Fig. 1 [2]. The input patterns are shown in Fig. 1(a), and desired clusters are shown in Fig. 1(b). Here points belonging to same cluster are given same label [2].

Fig. 1 Data Clustering [2]

Clustering, in data mining, is useful for discovering patterns. It helps in identifying interesting distributions inside the data [4]. Data clustering identifies the sparse and the crowded places. In this way it discovers the overall distribution patterns of the dataset [5]. Clustering techniques apply when there is no class to be predicted. The instances are divided into natural groups. A mechanism causes some instances to bear a strong resemblance to each other than they do to the remaining instances [6].

The remainder of this paper is organized as follows. In section 2, we give a brief description of data clustering methods. In section 3, we give a brief overview of CLUTO [11] clustering toolkit used in our experiments. In section 4, we give detail results of experiments. Section 5 contains review of the related work in the field. As for conclusions, they are provided in section 6.

II. DATA CLUSTERING METHODS

Clustering techniques are broadly divided into hierarchical clustering and partitional clustering.

A. Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity “[7].

Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [Jain and Dubes 1988; Kaufman and Rousseeuw 1990]. An agglomerative clustering starts with one-point (singleton)
clusters. It then recursively merges two or more most appropriate clusters [7]. Fig. 2 [8] provides a simple example of hierarchical clustering.

![Hierarchical Clustering](image)

Fig. 2 Hierarchical Clustering [8]

B. Partitional Clustering

Instead of a clustering structure a partitional clustering algorithm obtains a single partition of the data [2]. They generate a single partition of the data to recover natural groups present in the data. The proximity matrix among the objects is required by hierarchical clustering techniques. The partitional techniques expect data in the form of a pattern matrix. Partitioning techniques are used frequently in engineering applications where single partitions are important. Partitional clustering methods are especially appropriate for the efficient representation and compression of large databases [10].

The algorithm is typically run multiple times with different starting states. The best configuration obtained from all the runs is used as the output clustering [2]. Fig. 3 [4] provides a simple example of partitional clustering.

![Splitting of a large cluster by Partitional Algorithm](image)

Fig. 3 Splitting of a large cluster by Partitional Algorithm [4]

Dubes and Jain (1976) emphasize the distinction between clustering methods and clustering algorithms. The K-means is the simplest and most commonly used algorithm employing a squared error criterion [McQueen 1967]. The K-means algorithm is popular because it is easy to implement. Its time complexity is \( O(n) \), where \( n \) is the number of partitions. The algorithm is sensitive to the selection of initial partition. It may converge to a local minimum of the criterion function value if the initial partition is not properly chosen [2].

III. A CLUSTERING TOOLKIT: CLUTO

CLUTO [11] is a software package for clustering low- and high-dimensional datasets. It is used for analyzing the characteristics of the various clusters. CLUTO is well-suited for clustering data sets arising in many diverse application areas including information retrieval, web, science, and biology [11].

CLUTO provides three different classes of clustering algorithms. These algorithms operate either directly in the object’s feature space or in the object’s similarity space. These algorithms are based on the partitional, agglomerative, and graph-partitioning [16] paradigms. CLUTO provides a total of seven different criterion functions that can be used to drive both partitional and agglomerative clustering algorithms, which are described and analyzed in [12, 13].

gCLUTO is a cross-platform graphical application for clustering low and high-dimensional datasets. It is also used for analyzing the characteristics of the various clusters. gCLUTO is build on-top of the CLUTO clustering library [14]. wCLUTO is a web-enabled data clustering application that is designed for the clustering and data-analysis. wCLUTO is also built on top of the CLUTO clustering library [15].

IV. EXPERIMENTS AND RESULTS

Originally the Medview database has more than 8000 patient’s examinations and over 180 different attributes. But only those examinations have been considered that have values for the attributes “Diag-Def” and “Vis-cause=Primärundersökning”. The value of Viscause, “Primärundersökning”, corresponds to primary visits. Diag-Def attribute corresponds to definitive diagnosis. These two attributes are significant and plays vital role in classification. The set of attributes have been reduced and only those have been considered that are asked in common practice. The attributes are:

- Adv-drug
- Alcohol
- Allergy
- Bleed
- Care-provider
- Careprovider-now
- Civ-stat
- Diag-def
- Diag-hist
- Diag-tent
- Dis-now
- Dis-past
- Drug
- Family
- Health
- Lesn-on
- Lesn-site
- Lesn-trigg
- Mucos-attr
- Mucos-colr
- Mucos-site
- Mucos-size
- Mucos-txtur
- Ref-cause
- Smoke
- Snuff
- Symp-now
- Symp-on
The objective is to explore data clustering techniques for finding examinations that are similar to each other in the Medview dataset. In this regard, a series of experiments have been run with several clustering criterion functions for a selected algorithm. The algorithm has been run for a specified number of clusters. The clusters obtained are then evaluated to find an optimal number of clusters. The basic idea has been to investigate if a particular combination gives consistently good results. The number of clusters is then increased and the process is repeated until consistent results have been obtained with specified number of clusters.

Thus, there is a need to come up with a measure of cluster quality. CLUTO provides the clustering quality for each cluster as a function of its internal similarity and/or external similarity. It includes the average pair wise similarity between each object of each cluster and its SD and the average similarity between the objects of each cluster to the objects in the other clusters and their SD. The external similarity of an object is its similarity with objects in other clusters. Objects that have large values of internal similarity and small values of the external similarity tend to form the core of their clusters. A higher value of the internal similarity denotes a highly cohesive group.

A good solution should contain a number of clusters for which the sum of internal similarities of all the clusters is better and have a lower value of external similarity between objects of different clusters. This indicates the formation of highly cohesive groups. There is a possibility that the results are biased towards extremely small clusters. To negate this possibility, the number of objects in a cluster has to be included in the consideration.

Agglomerative and direct clustering techniques have not been so effective on our dataset. Where as Repeated Bisection technique have produced good results but Graph partitioning based method have produced far better results from the other methods. Fig. 4 shows a clustering solution obtained by applying Agglomerative clustering technique on the dataset with number of clusters = 30. In the Fig. 4, Cluster with id 0 has 860 objects in it. The cluster with id 1 has only 1 object.

The first column in Fig. 4 corresponds to cluster number (or cluster id). The second column shows number of objects in each cluster. The third column shows the average similarity between the objects of each cluster (i.e. internal similarities). The fourth column displays the standard deviation of these average internal similarities (i.e. internal standard deviations). The fifth column corresponds to average similarity of objects of each cluster and the rest of objects (i.e. external similarities). Finally, the last column corresponds to standard deviation of external similarities (i.e. external standard deviations).

Fig. 4 A Clustering Solution Obtained by Applying Agglomerative Clustering Technique

The first column in Fig. 4 corresponds to cluster number (or cluster id). The second column shows number of objects in each cluster. The third column shows the average similarity between the objects of each cluster (i.e. internal similarities). The fourth column displays the standard deviation of these average internal similarities (i.e. internal standard deviations). The fifth column corresponds to average similarity of objects of each cluster and the rest of objects (i.e. external similarities). Finally, the last column corresponds to standard deviation of external similarities (i.e. external standard deviations). Fig. 5 shows results obtained by applying Graph partitioning with Asymmetric graph model.
As an example, in Fig. 5 objects from two clusters have been described. The objects belonging to Cluster 0 tends to have more “Allergy” values as compared to objects belonging to Cluster 1. The “Allergy” values in cluster 0 have been over average. Similarly there has been a relationship between “vas-now” and “Symp-now”. It reflects to the fact that generally patients don’t complain if there is no symptom.

There has been a relationship between “Alcohol” and “Smoke” values. There has been a strange relationship between “Adv-drug” and “Care-provider”. “Adv-drug” is about the adverse effects the drugs have produced. This means that some dentists report more adverse drug reactions than the others. “Symp-site” (which is what patients normally tell about the symptoms) and “Mucos-site” (which is what dentists have to say about the symptoms) have a direct relationship.

There has been a strong correlation between Symptoms and Allergy values. Cluster 0 has given us a clue that patients with high allergy values are likely to have symptoms. Where as Cluster 1 comprises of examinations with no allergy and thus having no symptoms. This has been further visualized in Fig. 6 and Fig. 7. Fig. 6 shows objects in cluster 0 and this cluster contains examinations with symptoms. Fig. 7 shows objects in cluster 1 and this cluster contains patient examinations with no symptoms.

V. RELATED WORK

Medview [1] was designed earlier to support the learning process in oral medicine and oral pathology. The purpose of Medview was to provide a computerized teaching aid in oral medicine and oral pathology. In this regard, a clinical database was created from the referrals and has a large variation of clinical cases displayed by images and test based information. The students reach the database through the internet or other media. They can practice and learn at any convenient time. MedView contains search tools to explore the database and the students can study single cases or analyze various clinical parameters [1]. mEduWeb [1] is a web-based educational tool that allows students to search in the database and generate exercises with pictures of real patients [1]. mEduWebII was intended to enhance and make mEduWeb program better. It uses the MedView database containing
several thousand patient examinations [1]. Our work explored the possibilities of using Data Clustering techniques in oral medicine.

VI. CONCLUSION

Traditional data clustering techniques have been applied to find examinations that are similar to each other in the Medview dataset. In order to apply various clustering techniques, the first step has been the selection of a good algorithm. Each of clustering algorithms in the CLUTO package was applied on the dataset, and the results generated by all those experiments were studied in detail. Graph based partitioning clustering technique performed quite well although Repeated Bisection clustering algorithm also performed well. But Graph based clustering method produced balanced clusters. Balanced clustering has recently attracted an increased research interest. It has a good regularizing effect and decrease sensitivity to initialization. The results obtained have been useful while giving much information about the hidden patterns in the dataset. Moreover the results have been inspected by the clinician.

ACKNOWLEDGMENT

We would also like to thank Dr. George Karypis, from the University of Minnesota, Twin Cities, and others for having contributed the CLUTO clustering package to the research community.

REFERENCES