Blended Clustering for Health Data Mining

Arshad Muhammad Mehar, Anthony Maeder, Kenan Matawie, Athula Ginige

School of Computing & Mathematics, University of Western Sydney,
Locked Bag 1797 Penrith South DC, NSW 1797 Australia
{a.muhammad, a.maeder, k.matawie, a.ginige}@uws.edu.au

Abstract. Exploratory data analysis using data mining techniques is becoming more popular for investigating subtle relationships in health data, for which direct data collection trials would not be possible. Health data mining involving clustering for large complex data sets in such cases is often limited by insufficient key indicative variables. When a conventional clustering technique is then applied, the results may be too imprecise, or may be inappropriately clustered according to expectations. This paper suggests an approach which can offer greater range of choice for generating potential clusters of interest, from which a better outcome might in turn be obtained by aggregating the results. An example use case based on health services utilization characterization according to socio-demographic background is discussed and the blended clustering approach being taken for it is described.

Key words: data mining, data clustering, health data, health services utilization.

1 Introduction

Exploratory data analysis using data mining techniques is becoming more popular for investigating subtle relationships in health data, for which direct data collection trials would not be possible. Some examples of such problems include data collected for operational health service purposes (such as hospital admissions, or GP consultations, over a lengthy period of time), and population health data (such as longitudinal surveys of health status, or disease management and outcomes records).

Health data mining involving clustering for large complex data sets in these cases is often limited by insufficient richness in key decision variables. For example, in seeking to establish whether there is a difference in preventative health program outcomes for patients with different dietary habits, in the absence of explicit nutritional data, typically surrogate variables such as socio-economic status or even geographical location will be used as the indicative quantities.

A major deficiency with this approach is that the surrogate variables may be poorly or unstably correlated with the health outcomes under study. Consequently when a conventional clustering method is applied, the results may be too imprecise, or may be inappropriately clustered according to reasonable expectations. In the above example, if a standard method such as k-means clustering was applied to distinguish 3 groups (of poor, intermediate and good outcomes), it is possible that each of the clusters would include data points from across the whole range of surrogate variable values.

This paper suggests an approach which can offer greater range of choice for generating potential clusters of interest, from which a better data analysis outcomes might in turn be obtained by aggregating the results. By systematically applying a given clustering technique repeatedly, with different control parameters, a related succession of different clusters can be obtained, and the structure of these can be combined to yield regions in the data space for which greater strength of cluster membership may be inferred.

2 Health Services Utilization Studies for Large Data Sets

Health data research makes use of many different statistical and data mining techniques, to improve and maintain our health system and processes. As storage density increases and cost decreases exponentially, more and more transactional data is being collected in health. For example there are 5.7 million hospital admissions, over 200 million visits to doctors, and a similar number of prescribed medicines dispensed, that are captured electronically annually in Australia [1]. Analysis of this type of data offers benefits in many different areas such as health administration, adverse events, drug safety, disease diagnosis, population health and epidemiology. Often such analyses focus on well established performance indicators or quality metrics, such as length of stay or mortality. However, considerably more subtle structure is present in the data but is not easily extracted by simple clustering approaches. Some examples of attempts to find such structure in our area of interest of health services utilization characterization (specifically for hospital services) according to socio-demographic background are described below.

A study by [2] to investigate and compare hospital utilization in Victoria among Australian born and 8 different refugee source countries, showed that people born in refuge countries have lower or similar rates of hospitalization compared with Australian born. A random sample of 100,000 admissions of Australian born patients between 1998 and 2004 was compared with the total number (49,835) of admissions from non-Australian born patients from the source countries. Similar research in 1997 by [3] to determine the effect of hospital utilization between Danish born patients and non-Danish born immigrant patients showed that for certain types of cases Danish born patients consistently stay longer in hospitals than immigrants, and vice versa for other types of cases. This study included 5,310 persons discharged as inpatients, outpatients or emergency room patients who were born outside the Nordic countries, compared with a random sample of 10,000 patients born in Denmark. Another study on emergency hospital services (EHS) utilization in Spain by [4] was carried out to examine differences between immigrant and Spanish born people. This data set included patients between 15 and 64 years old for 96,916 hospital visits during the years 2004 to 2005. The results showed that people born outside Spain use EHS differently and more frequently than native born people in Spain. In Portugal an investigation of hospital utilization [5] collected data for 1,513 migrants. This work showed that age, length of stay, legal status and economic situation were interrelated in health services usage.

Narrower studies than the above have also been conducted, where specific types of disorder are of interest. An Australian study by [6] examined people with mental health problems who frequently attend the emergency department (ED) in tertiary referral metropolitan hospitals. The data was collected for 12 consecutive months between 2002 to 2003 year for 45,671 patients, from which 869 psychiatric patients and 1,076 presentations of these patients were identified. Significant differences were found for different age and diagnosis categories, for example younger people appeared more prominently in the frequent presenters group and also contained more anxiety/mood diagnosis than other groups. Another study in Spain was undertaken by [7] for 11,578 tertiary hospital admissions to psychiatric emergency services. Data collected included socio-demographic and clinical information that could be used to identify the difference between homeless and non-homeless patients and their admission patterns.

3 Data Mining Techniques for Health Data

To investigate patterns within large volumes of complex health data, some well established statistical methodologies are usually applied, based on hypothesis testing. In the past these efforts were limited primarily to epidemiological studies on clinical administrative and claims databases, due to lack of richness of other information [8]. Analogous problems to determine subtle patterns in patient cohort medical data have recently gained more attention. These problems have been addressed by applying techniques from machine learning and pattern recognition fields collectively, referred to as data mining [9].

The use of the term data mining originated from statistical computer science and is typically used in the context of large datasets [10]. It is a new generation approach to data analysis and knowledge discovery which has grown rapidly out of the need to derive useful knowledge from massive amounts of high dimensional data. Data mining is seen as exploratory rather than confirmatory in its approach [11]: a process of analyzing data from different perspectives and summarizing it into useful information (and hence also known as knowledge discovery [12]). Technically, it is the process of finding correlations or patterns among multiple data fields in large databases, using methods for searching through the data for patterns. Data mining leading to the extraction of hidden predictive information from large databases can help companies and organizations to focus on extracting important information to optimise their operations.

Data mining techniques are categorized into two different approaches: directed (or supervised) and undirected (or unsupervised). Supervised data mining is used mostly for hypothesis testing or verification, while unsupervised data mining is used mostly for new knowledge discovery [13]. Classification, estimation, and prediction are example of supervised data mining. Association rules, clustering and feature extraction are examples of unsupervised data mining. In this work we will take a simplified approach to demonstrate the principle we wish to describe, and so will concentrate on the use of clustering methods in unsupervised data mining.

Cluster analysis is the process of grouping a set of physical or abstract objects into classes according to some measure of similarity of the objects. It results in grouping of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [14]. For example, in medicine, clustering of symptoms of diseases can lead to very useful taxonomies for diagnosis. In psychiatry studies to find better therapies, correct diagnosis depends on clusters of symptoms to distinguish closely related disorders such as paranoia, schizophrenia, etc. Some commonly used clustering techniques which have been applied to health data mining are discussed below.

K-means clustering is a technique that separates a given set of data into k number of clusters (represented by centroids) in such a way that objects within a cluster are closer to their centroid than to the centroids of any other clusters [15]. Given n number of objects we choose arbitrarily k initial centroids where k is the number of desired clusters specified by the user. Each point is assigned to the closest centroid and each collection of points assigned to a centroid is a cluster. For each cluster the centroid position is updated based on the points assigned to the cluster, and this assigning and updating process is repeated until convergence or termination.

Hierarchical clustering techniques produce a structure of multi-scale hierarchical clusters, from small to large [16]. Agglomerative and divisive clustering are two different types of hierarchical clustering algorithms [17]. Agglomerative hierarchical clustering is constructed in a bottom up fashion in such a way that each data point is initially assigned to its own cluster, and these initial clusters are then successively combined according to their proximity. Divisive clustering takes a top-down approach, beginning with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until each object is in only one cluster, or until some termination condition holds.

Density based clustering techniques discern clusters of arbitrary shape in the database using models for data density and noise [18]. This approach determines height density regions for objects in data space, separated by a region of low density. The algorithm for density based clustering defines core points, border points and noise points. It has two input parameters MinPts and Eps, where MinPts is the minimum number of data points in any cluster and Eps is the threshold or maximum radius of cluster. Any two core points which are close enough to each other are assigned to the same cluster. Similarly, any border point which is close enough to a core point is assigned to the same cluster, while any noise points are eliminated [19].

4 Blended Clustering Approach

Data clustering techniques typically used in health data mining as identified above are generally based on fixed parameter choices, as there tend to be certain intuitive expectations about the number or nature of the clusters sought. For example, the number of clusters may be predetermined, or the variables used for the clustering may be selected from the overall set of variables in a data element. The nett effect of fixed parameter choices is that clustering results are usually obtained without any indication (or consideration) of sensitivity to the range of possible parameter choices. This means that the clusters obtained may be accepted as strongly formed, whereas in fact they may be arbitrarily dependent on the parameter choices. It also provides a result which defines cluster membership (or boundaries) in absolute terms, whereas considerable variations in the strength of membership may exist for elements which are quite close in the data space, or by the distance metric. These matters are exacerbated when the data sets are highly complex with subtle interdependencies, or when they have few variables that provide strongly correlated relationships with the nature of the clusters that are sought.

Our approach to overcoming these limitations is based on multiple applications of the clustering technique, and then combining the results of these applications to gain insights on the issues above. This approach requires different aspects of variation to be addressed, and in this paper we will consider only one such aspect, viz. the construction of *different numbers of clusters*. At least adjacent consecutive numbers, or a range around a value regarded as reasonable number (i.e. at least marginally more and less than the intuitively expected number of clusters).

Formally, we may describe our approach as follows. For a given choice of k = number of clusters, a given choice of clustering technique U, and a given choice of V = set of parameters $v_1..v_n$ used to control the clustering technique, we first construct a set of clusters $C_k(U,V) = \{c_{k,i}\}$ with i=1..k. Next, we construct sets of clusters $C_{k-1}(U,V)$, and $C_{k+1}(U,V)$ using the same clustering technique. In the work reported here, we will not vary U and V so we may write these cluster sets more simply as C_{k-1}, C_k and C_{k+1}.

In general, if k is somewhat less than the overall number of items being clustered, many items will be common between pairs of clusters across these three sets. If the items in common are fairly uniformly spread across all the clusters (i.e. changing the number of clusters has the effect of slightly increasing or decreasing membership of any given cluster, implying that clusters are of comparable strength) we have what we will term a "stable" clustering. On the other hand, if we find that items in common are non-uniformly spread across a few clusters (i.e. some clusters split fairly evenly into two or more clusters while other clusters are only slightly changed, implying that some clusters are of much greater strength than others), we will term this an "unstable" clustering.

Clustering is sensitive to choice of the related number of clusters and issues such as labeling, for the given data set. Due to this sensitivity, evaluation of the optimality of clustering is a common challenge. Stability is an indication of whether the algorithm fits the data points into the clusters strongly or not. Many validation measures such as Purity, Normalized Mutual Information, Rand index and F measure etc. are applied by [20] to evaluate the stability of clusters. A theoretical analysis on evaluation and stability was carried out by [21],[22] without any application to real datasets. Another study by [23] for evaluation of stability for k-means cluster ensembles with respect to random initialization, by using pairwise and nonpairwise methods. This study used only small and artificial data sets.

We can compute stability properties for the clusterings as follows. First we compare each of the k clusters in C_k with all of the k-1 clusters in C_{k-1}, and compare each of the k clusters in C_k with all of the k+1 clusters in C_{k+1}, to assess the proportion of data elements which are in common in each case. We denote the proportion of data elements in common between a particular pair of clusters, say cluster $c_{k,i}$ from C_k and cluster $c_{k-1,j}$ from C_{k-1} by $p(c_{k,i}, c_{k-1,j})$, which can be abbreviated to $p_{k,k-1,i,j}$. The set of maximum values of $p_{k,k-1,i,j}$ for each cluster $c_{k,i}$ in C_k compared with all the clusters in C_{k-1} is denoted by $P_{k,k-1}$. Similarly, we can compute $P_{k,k+1}$. Note that in general $P_{k,k-1}$ is not equal to $P_{k-1,k}$ as they have different cardinality. If each element of $P_{k,k-1}$ is above a given threshold T (i.e. for every cluster of C_k there is at least one cluster in C_{k-1} which has >T fraction of elements in common), we denote the clusterings as $T\text{-}_1$ stable for k; similarly we define T_{+1} stability for k. If the clusterings are both T_{-1} and T_{+1} stable as above, we call them T_1 stable for k.

T_1-stability indicates that the k-1, k or k+1 clusterings may all be considered to be valid options for interpreting the data, subject to the commonality constraint of threshold T. In this case we may wish to proceed with further analysis by tightening the value of T, or by constructing sets C_{k-2} and C_{k+2} to assess their T_2 stability. If either T_{-1} or T_{+1} stability holds but not the other, C_k must represent a point at which some degree of change or discontinuity occurs for structural properties in the data set. If neither T_{-1} nor T_{+1} stability holds, we may wish to consider whether another value for T may be more appropriate.

5 Sample Experimental Results

Consider the example of investigating different patterns of hospital service utilization for a highly heterogeneous population which has a wide mix of age, ethnicity and socio-economic status factors. In a typical hospital admissions dataset, very few variables directly indicative of these factors are available, and therefore surrogates such as country of birth or home language must be used. The occurrence and number of distinct classes for different utilization patterns is unknown a priori. The above stability analysis approach thus offers one way to evaluate plausible options for both number and membership of classes, and choices of variable values to achieve these.

Table 1: Sample data set of 20 patients from 4 countries, varied by severity and age.

Patient#	1	2	3	4	5	6	7	8	9	10
Country	A	A	A	A	A	B	B	B	B	B
Severity	9	5	4	9	7	5	8	4	7	8
Age	55	45	65	25	50	20	25	35	45	50

Patient#	11	12	13	14	15	16	17	18	19	20
Country	B	C	C	C	C	D	D	A	D	D
Severity	9	7	6	4	3	5	4	7	3	7
Age	55	65	55	70	25	20	60	20	65	20

To explore this approach, we constructed a sample data set of 20 elements with properties similar to population characteristics for residents of Greater Western Sydney (see Table 1). This data set contains a range of patient profiles from 4 different ethnic origins, aged from 20 to 70 years old, and with disease severities coded on a 10 point scale.

We made a single, randomly seeded application of k-means clustering, based on joint age and severity variables with k=2, 3, 4. The results of this clustering (see Table 2) were used to determine that at the level T=0.5, $C_{3,1}$ and $C_{3,3}$ are T_1-stable. By inspection, it was found that $C_{3,1}$ comprised membership of 50% from country A, while $C_{3,3}$ comprised membership of 50% from country B. This allowed us to nominate k=3 as a closer match with the data than either k=2 or k=4, and furthermore to infer that subjects of ethnic origin from countries A and B were more self-similar in the relationship between their age and disease severity characteristics, than those of countries C and D.

Table 2: Results of k=2, 3, 4 k-means clusterings for the data of Table 1.

Cluster	C2,1	C2,2	C3,1	C3,2	C3,3	C4,1	C4,2	C4,3	C4,4
Members	1	4	1	3	4	1	2	3	4
	2	8	2	12	6	5	7	8	9
	3	9	5	14	7	6	13	14	10
	5	10	9	17	8	12	17	19	11
	6	11	10	19	15	16			15
	7	15	11	20	16				18
	12	17	13		18				20
	13	18							
	14	19							
	16	20							

The above approach does not take into account some other aspects of data clustering techniques which might also affect the results, such as choice of different clustering algorithms, choice of different cluster membership (or distance) functions, choice of different control parameters to control the clustering algorithm, or choice of different subsets of variables describing items in the data set to influence the nature of the clusters. While these aspects may in principle be treated in a similar manner, they are substantially more complex in the nature of the choice, and likely to have far stronger effect than the number of clusters choice. Blending involving these aspects is intrinsically more expensive and is likely to require more careful control than the stability-based strategy described above. Consequently they have not been considered further here.

6 Conclusion

The work described here offers a first step towards what may be seen as a blended form of an "adaptive" or "learning" approach to clustering, which identifies patterns emerging from repeated application of "perturbed" configurations of the chosen data mining method based on the simplest control aspect of clustering, viz. the number of clusters. In some sense this approach is closer in philosophy to a genetic algorithm technique than a rule-based technique, as preconditions for the allowable configurations are much looser than would be determined by a hypothesis driven conception. In order to gain greater efficiency, it would be desirable to limit the range over which perturbed configurations can occur. It would also be beneficial to develop more structure in the approach to inform the choice of perturbations, and the ways of combining them to reach the improved conclusions about the data.

References

1. McAullay, D., et al.: A delivery framework for health data mining and analytics. Australian Computer Society Inc., Darlinghurst (2005)
2. Correa-Velez, I., et al.: Hospital utilisation among people born in refugee-source countries: An analysis of hospital admissions, Victoria, 1998-2004. Medical Journal of Australia,. **186**(11): 577 (2007).
3. Krasnik, A., et al.: Effect of ethnic background on Danish hospital utilisation patterns. Social Science & Medicine, **55**(7), 1207-1211 (2002)
4. RuÃc, M., et al.: Emergency hospital services utilization in Lleida (Spain): A cross-sectional study of immigrant and Spanish-born populations. BMC Health Services Research, **8**(1), 81-90 (2008)
5. Dias, S.n.F., Severo M., Barros H.: Determinants of health care utilization by immigrants in Portugal. BMC Health Services Research, **8,** 1-8 (2008)

6 Brunero, S., et al.: Clinical characteristics of people with mental health problems who frequently attend an Australian emergency department. Australian Health Review, **31**(3), 462-470 (2007)

7. Pascual, J.C., et al.: Utilization of psychiatric emergency services by homeless persons in Spain. General Hospital Psychiatry, **30**(1), 14-19 (2007)
8. Prather, J., et al.: Medical data mining: knowledge discovery in a clinical data warehouse. American Medical Informatics Association (1997)
9. Harrison, J.H., Jr.: Introduction to the mining of clinical data. Clinics in Laboratory Medicine, **28**(1), 1-7 (2008)
10. Dominique, H., et al.: A review of software packages for data mining. The American Statistician, **57**(4), 290 (2003)
11. Tukey, J.: Exploratory data analysis. Addison-Wesley, Reading (1977)
12. Fayyad, U.M.: Advances in knowledge discovery and data mining. AAAI Press : MIT Press, Menlo Park (1996)
13. Berger, A.M.M.M.R. and C.R.M.M. Berger: Data Mining as a Tool for Research and Knowledge Development in Nursing. CIN: Computers, Informatics, Nursing, **22**(3), 123-131 (2004)
14. Han, J. and M. Kamber: Data mining : concepts and techniques. Morgan Kaufmann series in data management systems. Morgan Kaufmann, San Francisco (2006)
15. Mwasiagi, J., X. Wang, and X. Huang: The use of k-means and artificial neural network to classify cotton lint. Fibers and Polymers, **10**(3), 379-383 (2009)

16.Hair, J.F.: Multivariate data analysis. Prentice Hall, Upper Saddle River (1998)

17.Crowley, J. and D. Ankerst: Handbook of statistics in clinical oncology. CRC Press (2006)

18.Berry, M.W. and M. Brown: Lecture notes in data mining. World Scientific, Hackensack (2006)

19.Tan, P.-N., V. Kumar, and M. Steinbach: Introduction to data mining. Pearson Addison Wesley, Boston (2005)

20.Wu, J., H. Xiong, and J. Chen: Adapting the right measures for k-means clustering, in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining., ACM: Paris, France., 877-886 (2009)

21.Ben-David, S., et al.: Stability of k-means clustering, in Proceedings of the 20th Annual Conference on Learning Theory., Springer-Verlag: San Diego, CA, USA., 20-34 (2007).

22.Rakhlin, A. and A. Caponnetto: Stability of k-means clustering. Advances in Neural Information Processing Systems,. **19**: 1121-1127(2007).

23.Kuncheva, L.I. and D.P. Vetrov: Evaluation of Stability of k-means Cluster Ensembles with Respect to Random Initialization. Pattern Analysis and Machine Intelligence, IEEE Transactions on,. **28**(11): 1798-1808 (2006).