IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4187

Applying Data Mining Techniques To Improve Data Quality of Patient Records

Narasimhaiah Gorla
Associate Profesor of Information Systems, Wayne State University, Michigan
Tel: (313) 577-2568, Fax: (313) 577-4880, n_gorla@wayne.edu

Chow Y. K. Bennon
Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

Public hospitals are under the control and supervision of the Hospital Authority in Hong Kong. The demographic and clinical description of each patient is recorded in the databases of various hospital information systems. The errors in patient data result in erroneous conclusions by the doctors and lost time to resolve data errors. The reason for data errors are wrong entry of data, absence of information provided by the patient when they enter the hospital, improper identity of the patients (especially in case of tourists) etc. All these factors will lead to a phenomenon that several records of the same patient will be shown as records of different patients.

In this research, we illustrate the use of "clustering" technique, a data mining technique, the hospital can use to group "similar" patients together. We use two algorithms: hierarchical clustering and partitioned clustering. Furthermore, we combined these two algorithms to generate "hybrid" clustering algorithm and applied on the patient data, using a C program. We used six attributes of patient data: Sex, DOB, Name, Marital Status, District, and Telephone number as the basis for similarity of patient records. We also used some weights to these variables in computing similarity. We found that the Hybrid algorithm gave more accurate grouping compared to the other algorithms, had smaller mean square error, and executed faster. Due to the privacy ordinance, the true data of patients will not be shown, but only simulated data will be used.

INTRODUCTION

The establishment of the Hospital Authority marked the milestone for all public hospitals in Hong Kong. Under the management of the Hospital Authority, the daily operations of the public hospitals are linked and are under the centralized control of Hospital Authority Head Office (HAHO). Whenever a person using any service of any one of the public hospitals, his / her personal information will be entered into a corresponding clinical information system. All the clinical information systems are linked and share a master database system, which is known as the Patient Master Index (PMI). The main function of PMI is to maintain and control the demographic data of patients. Each patient record is identified by his / her Hong Kong Identification Card (HKID) number or Birth Certificate number. The personal data in PMI can be accessed by any linked clinical information system. Whenever the detail of a patient is updated by any linked clinical information system, PMI is updated automatically. Ideally, each patient should have a unique master record in PMI. However, a patient may contain two or more records in the PMI due to reasons, such as, a patient comes to a hospital without his / her HKID, the information of the patient is wrongly reported by a second party; a patient cannot be identified by any valid document, the patient is a non-Hong Kong resident such as tourist, a baby is born without a Birth Certificate, data entry operators make typing errors.

If any of the above cases occur, a unique pseudo-identification number will be generated by the clinical information system to identify the patient temporarily. Once the HKID or Birth Certificate number is provided by the patient during hospitalization, the personal information will be updated and while the pseudo-identification number will be discarded. However, many patients cannot provide their identification documents before they leave the hospital. More temporary records may be generated for the same patient when a patient attends a hospital several times, thus more and more "duplicate" records will be generated in the PMI database.

Objective of Study

The objective is to group records in the PMI database based on the similarity of attributes of the patients. Those records with high similarity, presumably belonging to the same patient, will be clustered and reported. Hospital staff can make use of the reports to study the identity of "similar" records and merge them as necessary. The benefits of this study to the hospital are: i) time can be reduced on producing accurate patient records, ii) searching for data can be speeded up since dummy records are reduced iii) the historical data of a patient becomes more complete and accurate, and iv) the quality of patient records is improved, in general.

Literature Review

Cluster analysis is the formal study of algorithms and methods for grouping, or classifying similar objects, and has been practiced for many years. Cluster analysis is one component of exploratory data analysis (EDA) (Gordon, 1981). Cluster analysis is a modern statistical method of partitioning an observed sample population into disjoint or overlapping homogeneous classes. This classification may help to optimize a functional process [Willett, 1987]. A cluster is comprised of a number of similar objects collected or grouped together. "A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it" [Everitt, 1974].

DATA MODEL

Selection of Patient Attributes

In this research, the goal is to group together the records with similar attributes. Large number of variables (attributes) is not desirable to find the similarity-reduction and selection of variables is a fundamental step in cluster analysis (Hand, 1981). In Systems, i.e., PMI database, a patient is characterized by the following variables:

a) HKID number / Birth Certificate number
b) Name in English
c) Name in Chinese
d) Marital status
e) Sex
f) Date of birth
g) Age
h) Address / District code
i) Telephone number
j) Nationality
k) Religion

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

The values of some of these variables may change considerably over time, or they are not filled in the database. In order to determine which variables to be chosen, three basic criteria are used: stability of the variable, accuracy of the variable, and importance of the variable. It is important to reduce number of variables, since computer time increases dramatically with an increase in the number of variables (Anderberg, 1973). Since the HKID numbers of two records belonging to the same patient are different or the HKIDs are missing, HKID is not a suitable variable for clustering purposes. The variables of nationality and religion are optional input to the hospital information system and we find that over 80% of patients do not provide such information-hence nationality and religion are not suitable. The variable age is not suitable, since patients born in different months of same year may be grouped into the same age. The variable name in Chinese is a good choice for clustering, but the program development in handling Chinese is more complicated, so we do not use it. So, we are left with six variables for computing similarity:

1. Sex
2. Name in English
3. Date of birth
4. Marital status
5. District code
6. Telephone number

Furthermore, marital status, district code and telephone number are unstable variables. They often change because patient can get married; change his/her residence and telephone number. However, these changes occur rarely, so the reliability of these variables remains quite high. The variable Name in English may also be changed, but it is rare. The final variable date of birth will not change. Thus, we use the above six variables and use different combinations of these variables to obtain good clusters.

DATA REPRESENTATION, STANDARDIZATION, AND WEIGHTING

Data Representation

Clustering algorithms group objects based on indices of proximity between pairs of objects. Each object is represented by a pattern or d-place vector, which indicates the d-attributes of the object. The collection of each d-place pattern forms a pattern matrix. Each row of this matrix defines a pattern and each column denotes a feature or attribute. The information of the pattern matrix can be used to build a proximity matrix. A proximity matrix [d(i,j)] gathers the pairwise indices of proximity in a matrix in which each row and column represents a pattern. A proximity index can represent either a similarity or dissimilarity. The more the i^{th} and j^{th} objects resemble one another, the larger a similarity index and the smaller a dissimilarity index. Apart from the number of patterns *n*, only the upper triangle of the proximity matrix is required as the proximity matrix is assumed to be symmetric. The properties of a proximity index between the i^{th} and k^{th} pattern [denoted as d(i,k)] have been summarized as:

1) (a) For a dissimilarity, d(i,i) = 0, for all i.
   (b) For a similarity, d(i,i) > = max d(i,k), for all i.
2) d(i,k) = d(k,i), for all (i,k).
3) d(i,k) > 0, for all (i,k).

A C-program is written to perform the cluster analysis. As stated earlier, six attributes of each patient will be used for our analysis: sex, date of birth, name, marital status, district code, and telephone number. The HKID number is used internally for testing purposes but not to compute clusters.

Dissimilarity

Comparison between two objects can be expressed in either similarity or dissimilarity. The Euclidean distance, a mathematical equation that measures the distance between two objects, is used for measurement. Let $[x_{ij}]$ be our pattern matrix, where x is the j^{th} feature for the i^{th} pattern. The i^{th} pattern, which is the i^{th} row of the pattern matrix, is denoted by the column vector x_i.

$$x_i = (x_{i1}, x_{i2}, \ldots\ldots\ldots x_{id})^T, \quad i = 1,2,\ldots\ldots,n$$

The Euclidean distance between object i and object k can then be expressed as:

$$d(i,k) = [\sum_{j=1}^{d} (x_{ij} - x_{kj})^2]^{1/2}$$

As our data file consists both ordinal and binary values, we handle it in the following way. Let attribute of a patient be j = {1,2,3,4,5,6} equivalent to {sex, date of birth, name, marital status, district code, telephone number} respectively.

For any object i or k, if j = 1,4,5,6, (i.e., attribute = sex, marital status, district code or telephone number),

$$(x_{ij} - x_{kj}) = \begin{cases} 0, & \text{if } x_{ij} = x_{kj} \\ 1, & \text{otherwise} \end{cases}$$

If j = 2 (i.e., attribute = date of birth),

$(x_{ij} - x_{kj})$ = the number of days between two date of birth.

If j = 3 (i.e., attribute = name),

$(x_{ij} - x_{kj})$ = the fraction of matched characters between two names.

The sum of these six individual values, by the definition of Euclidean distance, represents the overall dissimilarity between two objects, which forms the basic component in our proximity matrix. For example, consider the following two rows from the pattern matrix,

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} M & 02071972 & \text{CHAN TAI MAN} & S & NP & 23121256 \\ F & 02081972 & \text{CHAN YEE MAN} & S & QRB & 27191928 \end{bmatrix}$$

for j=1, $(d_{11} - d_{21}) = 1$ (Not the same sex)
for j=2, $(d_{12} - d_{22}) = 31$ (No. of days between date of births)
for j=3, $(d_{13} - d_{23}) = 1 - 9/12 = 0.25$ (% of no. of mismatch characters)
for j=4, $(d_{14} - d_{24}) = 0$ (Both are single)
for j=5, $(d_{15} - d_{25}) = 1$ (Not the same district code)
for j=6, $(d_{16} - d_{26}) = 1$ (Different telephone no.)

By the definition of Euclidean distance, the dissimilarity between x_1 and x_2 is

$$d(1,2) = [1^2+31^2+0.25^2+0^2+1^2+1^2]^{1/2} = 31.0494$$

Missing Data

In practice, measurement pattern vectors can be incomplete because of errors, or unavailability of information. Jain [1988] described a simple and general techniques for handling such missing values. The distance between two vectors x_i and x_k containing missing values is computed as follows. First define d_j between the two patterns along the j^{th} feature:

$$d_j = \begin{cases} 0, & \text{if } x_{ij} \text{ or } x_{kj} \text{ is missing} \\ x_{ij} - x_{kj}, & \text{otherwise} \end{cases}$$

Then the distance between x_i and x_k is written as

$$d(i,k) = \left[\frac{d}{d - d_o} \sum d_j^2 \right]^{1/2}$$

where d_o is the number of features missing in x_i or x_k or both. If there are no missing values, then d(i,k) is the Euclidean distance. In this research, a missing data is represented by "0" in the data file. Now, suppose the telephone number of patient 2 is missing (i.e. $x_{26} = 0$) in the previous example,

$$d_0 = 1$$
$$\sum d_j^2 = 1^2+31^2+0.25^2+0^2+1^2+0^2 = 963.0625$$
$$d(1,2) = [6 / (6 - 1)] \times 963.0625 = 33.9952$$

Data Standardization

The individual dissimilarity value for all attributes of a patient falls between 0 and 1, except the attribute of date of birth, whose value varies from 0 to hundreds of thousands. This creates a bias in the cluster analysis towards date of birth. So normalization is performed so that the range of all measurements is 0 to 1. The standard normalization for a number, x_{ij}, can be expressed as:

$$x_{ij}^* = \frac{x_{ij} - m_j}{s_j}$$

where $$m_j = \left(\sum_{i=1}^{n} x_{ij} \right) / n$$

and $$s_j^2 = \left[\sum_{i=1}^{n} (x_{ij} - m_j)^2 \right] / n$$

(x_{ij} is the actual magnitude of the variable, m_j is the mean of all x_{ij} for attribute j, and s_j is the standard deviation of all x_{ij} for attribute j)

Consider the previous example again, let there are 100 patient records (i.e. n=100). x_{12} can be represented as the number of day between the date of birth and a standard date of 31121999. Therefore,

x_{12} = days difference between 02071972 and 31121999 = 10039

x_{22} = days difference between 02081972 and 31121999 = 10008

Now, suppose $m_2 = 18000$, $s_2 = 8000$, for 100 patient records,

$$x_{12}^* = (10039 - 18000) / 8000 = -0.9951$$

$$x_{22}^* = (10008 - 18000) / 8000 = -0.9999$$

$$(d_{12} - d_{22}) = (-0.9951) - (-0.9999) = 0.0039$$

$$d(1,2) = [1^2+0.0039^2+0.25^2+0^2+1^2+1^2]^{1/2} = 1.7569$$

Weighting

As discussed before, some attributes change more often than others, over a period of time. When doing the cluster analysis, we need to give more weight to more stable attributes. Different weightings will result in different clusters. Applying the weighting factors, the Euclidean distance between object i and object k becomes:

$$d(i,k) = \left[\sum_{j=1}^{d} w_j (x_{ij} - x_{kj})^2 \right]^{1/2}$$

CLUSTERING ALGORITRHMS

Hierarchical Clustering

A hierarchical clustering is a sequence of partitions in which each partitional is nested into the next partition in the sequence. The algorithm for hierarchical clustering starts with the disjoint clustering, which places each of the *n* objects in an individual cluster. Two methods for updating the proximity matrix are provided in the C program: Single-link and compete-link. The first Single-link clusters are based on connectedness and are characterized by minimum path length among all pairs of objects in the cluster. The proximity matrix is updated with $d[(k),(r,s)] = \min \{d[(k), (r)], d[(k),(s)]\}$, which corresponds to $\alpha_r=0.5$, $\alpha_s=0.5$, $\beta=0$ and $\gamma= -0.5$. Complete-link clusters are based on complete subgraphs where the diameter of a complete subgraph is the largest proximity among all proximities for pairs of objects in the subgraph.

The number of clusters formed depends on the value of a threshold of level. From the above graph, it can be seen that a small threshold will return a larger number of clusters, while a large threshold will return a smaller number of clusters. In this research, we would like to obtain the most similar patient records, which implies large number of clusters; thus it requires a smaller threshold value.

Partitional Clustering

A partitional clustering determines a partition of the *n* patterns into *k* groups or clusters such that the patterns in a cluster are more similar to each other than to patterns in other clusters. The value of *k* may or may not be specified. The most commonly used partitional clustering strategy is based on the square error criterion. The objective is to obtain the partitions, which minimizes the square error for a given number of clusters. Suppose that a given set of *n* patterns in *d* dimensions has been partitioned into *K* clusters $\{C_1, C_2, \ldots\ldots C_K\}$ such that cluster C_K has n_K patterns and each pattern is in exactly one cluster, so that

$$\sum_{i=1}^{K} n_i = n$$

Hybrid Clustering

In our research, we also use a hybrid clustering, in which the hierarchical and partitional algorithms are combined and modified to produce clustering. We compare the results of the hybrid algorithm with the other two clustering algorithms. From the output of hierarchical clustering, we can obtain different cluster combinations for a different threshold. They serve as the initialization for the partitional clustering. The hybrid clustering algorithm can be described as follows.

Step 1: Select an initial partition from one of the results of hierarchical clustering by adjusting different threshold of the level.

Step 2: Generate a new partition by assigning each pattern to other clusters one by one.

Step 3: Compute new cluster centres as the centroids of the clusters as well as the overall square error.

Step 4: Choose the partition with the lowest value of overall square error.

Step 5: Repeat Step 2 - 4 for other patterns.

Step 6: Repeat Step 2 - 5 until an optimum value of the criterion function is found or the specified number of iteration is reached.

RESULTS

In order to test the clustering algorithms, simulation studies involving the generation of artificial data set for which the true structure of the data is known in advance. The performance of the clustering method can then be assessed by determining the degree to which it can discover the true structure.

First Experiment

A simulated data set is used to compare the power of the three clustering algorithms. For an easy identification of duplicated records, 20 records are used in which 5 of them are with repetitions. The simulated input data is shown in Appendix 1. In hierarchical clustering, both the single-linkage method and the complete-linkage method are used to cluster the records. The corresponding dendrograms for both the single-linkage and complete-linkage are shown in Appendix 2. In both methods, the duplicated records are grouped but with different subsequent formation of clusters. From the dendrograms, it can be seen that a small threshold will produce the same clustering result.

The clustered output data by partitional clustering and hybrid clustering are shown in Appendix 3 and 4 respectively. The summarized results for the three algorithms are shown in Figure 1.

From Figure 1, it can be seen that the mean square error of hierarchical clustering is larger than that of partitional clustering. However, the percentage of error in hierarchical clustering is smaller than that of partitional clustering when the number of iteration is small. The result of partitional clustering can further be improved by increasing the number iteration so as to reduce the mean square error, and percentage of error. At the same time, it increases the time for convergence. Hybrid clustering combines the advantages of both from the hierarchical clustering and partitional clustering. With a good

Figure 1: The best results for the three clustering algorithms with 20 input records

Algorithm	No. of cluster	mean square error		% of error		No. of iteration
		iteration	iteration	iteration	iteration	for convergence
Hierarchical	5	1124	-	0	-	-
	10	60.4	-	0	-	-
	15	20	-	0	-	-
Partitional	5	101.9	91.55	40	40	>500000
	10	40.45	38.9	60	60	>500000
	15	33.4	30.25	100	80	>500000
Hybrid	5	227.05	201.55	0	0	<50000
	10	36.8	30.1	0	0	<50000
	15	20	20	0	0	<100

choice of threshold, records are well structured into clusters by the initializing hierarchical clustering. Then the structure is further re-organized in order to reduce the mean square error by the following partitional clustering. As the choice of initial number of cluster is first determined by the hierarchical part, number of iteration can highly be reduced to get the local minimum or even the global minimum. The rate of convergence is much faster than that of partitional clustering.

Second Experiment

A second simulation study is also performed but with a much larger data set. 100 records are used in which 10% are with repetition. The summarized results from the hierarchical clustering, partitional clustering and hybrid clustering are shown in Figure 2 to 4, respectively.

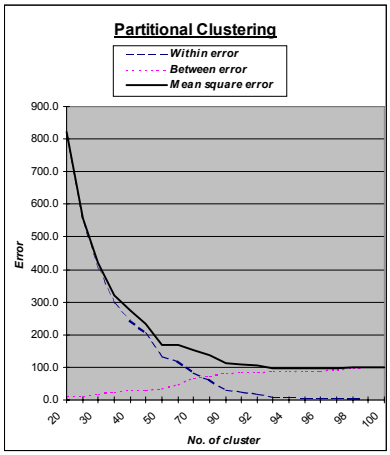
Figure 2: Summarized result for the hierarchical clustering with 100 input records

Threshold	No. of cluster	Mean square error	% of error
0.5	90	100	0
1	90	100	0
1.5	25	865	0
2	2	71249	0
3.5	2	72215	0

As can be seen in Figure 2, the best solution is obtained without any error by the hierarchical clustering (number of cluster=90). It forms the initialization of number of clusters in hybrid clustering. The structure of the result can further be reduced with a decrease in mean square error by increasing the number of iteration (number of cluster=88 in Figure 4). In Figure 3, the best solution is obtained when number of cluster is 93 with 10% of error. This solution is only the local minimum with 1000 iterations. A global minimum appears when the number of iteration is increased so that there is no error. The rate of convergence is much lower in partitional clustering than the hybrid clustering. As seen in Figure 4, the hybrid clustering converges after 500 iterations but the partitional clustering still oscillates.

Figure 3: Summarized result for the partitional clustering with 100 input records

No. of cluster	Within Error	Between Error	Mean sq. err
5	12006.2	3.8	12010.0
10	3075.9	8.1	3084.0
15	1411.8	11.8	1423.6
20	810.2	12.0	822.2
25	545.1	12.9	558.0
30	402.0	19.8	421.8
35	295.8	23.9	319.7
40	244.0	30.2	274.2
45	201.6	31.2	232.8
50	134.0	33.3	167.3
60	118.6	48.2	166.8
70	82.7	69.4	152.1
80	61.0	75.4	136.4
90	30.9	81.0	111.9
91	25.2	83.1	108.3
92	20.2	85.2	105.4
93	9.1	86.4	95.5
94	8.5	87.5	96.0
95	7.4	89.4	96.8
96	5.9	91.2	97.1
97	5.2	92.7	97.9
98	2.7	95.8	98.5
99	1.0	98.3	99.3
100	0.0	100.0	100.0



Comparison

The primary step in hierarchical clustering is to group those records with high similarity. Once the records are grouped into any cluster, they will not be further considered individually. The similarity of the whole cluster will then be compared with other records to form a much larger cluster, if necessary. On the other hand, partitional clustering considers all the records every time and tries to group them into different clusters with the total least difference among the elements within clusters. It needs a very long time to re-structure the clustering before a fairly good result can be obtained. The combination of the two algorithm forms the hybrid clustering algorithm. It tries to reduce the time of iteration required in partitional clustering by constructing a quite good starting structure using hierarchical clustering first. The objective is then achieved by further searching the global minimum value of the overall mean square error.

CONCLUSION

In this research, we illustrated the use of data mining techniques, particularly clustering algorithms, to identify duplicate patient records in Hospital Information System in Hong Kong. The reason for errors in patient data are wrong data entry, insufficient information provided by the patient, improper identity of the patients (especially in case of tourists in Hong Kong) etc. The clustering algorithms are used in this

Figure 4: Summarized result for the hybrid clustering with 100 input records

Algorithm	Threshold	No. of cluster	Mean square error				
			No. of iteration				
			0	100	200	500	1000
Hybrid	0.5	90	100	95.2	94.4 (reduce to 88 clusters)	94.4 (reduce to 88 clusters)	94.4 (reduce to 88 clusters)
	1	90	100	95.2	94.4 (reduce to 88 clusters)	94.4 (reduce to 88 clusters)	94.4 (reduce to 88 clusters)
	1.5	25	865	824	806	794	794
	2	2	71249	70867	70584	70151.9	70151.9
Partitional		90		132.01	132.01	132.01	111.88
		25		596.71	596.71	596.71	570.71
		2		72105.97	72105.97	72105.97	7274.97

research to cluster all the records of each patient together, so that high quality data records of each patient can be maintained. The importance of this research cannot be over-emphasized because of the fact that wrongly grouped records or ungrouped records of a patient will give raise to inaccurate or missing information about the patient medical history, which will further give raise to erroneous conclusions by the doctors. We used three clustering algorithms: hierarchical, partitioned, and hybrid that combines the other two. Our results show that Hybrid algorithm gave more accurate grouping compared to the other algorithms, and result was obtained faster (fewer iterations).

REFERENCES

Backer, E, 1995, "Computer-assisted Reasoning in Cluster Analysis". NY: Prentice Hall.

Sholom M. Weiss, 1991, "Computer Systems That Learn". California : Morgan Kaufmann.

E. Diday, 1994, "New approaches in classification and data analysis". NY : Springer-Verlag.

Everitt, Brian S, 1974, "Cluster Analysis". London: E. Arnold.

Hand, DJ, 1981, "Discrimination and Classification", NY: J. Wiley.

Willett, Peter, 1987, "Similarity and clustering in chemical information systems". NY : Wiley.

Hubert, L.J., 1995, "Clustering and classification". Singapore : World Scientific.

Anderberg, Michael R, 1973, "Cluster analysis for applications". NY : Academic Press

Modell, Martin E., 1992, "Data Analysis, Data Modeling, and Classification". NY: McGraw-Hill.

Rencher, Alvin C, 1995, "Methids of multivariate analysis". New York : Weley.

Jain, Anil K., 1988, "Algorithms for Clustering Data". NJ: Prentice Hall.

McLachlan, Geoffrey J, 1988, "Mixture models : inference and applications to clustering". NY : M. Dekker.

Gordon, AD, 1981, "Classification: Methods for the Exploratory Analysis of Multivariate Data". London: Chapman and Hall.

Spath, Helmuth, 1980, "Cluster analysis algorithms for data reduction and classification of objects". NY : Halsted Press.