# Unsupervised Learning Aided by Clustering and Local-Global Hierarchical Analysis in Knowledge Exploration

Yihao Zhang, Mehmet A. Orgun, Weiqiang Lin
Department of Computing Change Program
Macquarie University Australian Taxation Office
North Ryde, NSW 2109, Australia Sydney, NSW 2000
Australia
[{yihao, mehmet}@ics.mq.edu.au] [wei.lin@ato.gov.au]

Journal of Digital Information Management

***ABSTRACT:*** *Unsupervised learning plays an important role in the knowlede exploration discovery. The basic task of unsupervised learning is to find latent variablesor relationships in a given dataset wihout any assumed regularities or patterns. In this paper we apply two advanced models, clustering analysis and hierarchial analysis to accomplish unsupervised learning. K-Means clustering presents its strength in large scale clustering. The original data can be pre-processed and the potential variables are targeted. Correlations among these variables are explored in the subsequent sets by Local Global Hierarchial Analysis (LGHA) assisted by three main steps. In the first step, we use a structural approach to find qualititative patterns from the given variables. Then, the second step applies a quantitative based algorithm to find quantitative patterns from those variables. The and last step generated global hybrid patterns by combining the local patterns obtained from the first two steps based on a certain criterion. Both of the K-Means and Local Global Hierarchial Analysis (LGHA) models are applied in experiments with real world longitutional medical datasets.*

**Categories and Subject Descriptors**
**I 2.4 [Knowledge representation formalism and methods**]; **I.2.6** [**Learning**]: **I.5.3 [Clustering**]

**General Terms**
Knowlege representation, Learning classification

**Keywords**: Hierarchical Analysis, K-Means Clustering, Unsupervised Learning, Knowledge Exploration

Received 30 November 2006; Revised 15 February 2007; Accepted 12 March 2007

## 1. Introduction

From a traditional point of view, knowledge exploration can be categorized into supervised learning and unsupervised learning (Jordan and Jacobs 1994). In the last decade, there have been research activities on supervised learning approaches and techniques, whereby class information is available before any knowledge exploration takes place. The most utilized approach is to achieve a predetermined independent measurement in order to preferentially target classes. Then a classification algorithm is applied in the data pre-processing stage (Liu and Motoda 1998, Liu and Yu 2005). However, this approach is not robust to be effectively applied on features with irregular sizes or nonrecurring, high-dimensional variables.

Unsupervised learning is a recent approach in knowledge exploration. It is widely used on/with unlabeled data, such as extracting relevance that exists in records. Unsupervised learning is an important supplementary method to category data since it could increase the precision of clustering results. Unlike supervised learning, unsupervised learning attempts to find the most reasonable patterns by uncovering relationships best instead of using preferential classification labels (Dy and Brodley 2000, 2004). Because the idea behind unsupervised learning is to run an unsupervised algorithm on raw data (Kohavi and John 1997), most researchers consider the applications of data clustering and data reduction (including dimension reduction, size reduction, etc.) as two key issues in the framework of knowledge exploration. The use of an unsupervised learning method could save time in data processing by removing the matching and ranking process used for specified classes, and avoiding redundant analysis.

In this paper, we propose to combine two models to achieve unsupervised learning. K-Means Clustering Analysis (K-Means) is used to partition the original combine two models to achieve unsupervised learning. K-Means Clustering Analysis (K-Means) is used to partition the original data according to a certain criterion. As a robust model, K-Means semiautomatically generates clusters and assigns data into different clusters. The data within these clusters will be labelled prior to when we collect observational sets.

Local-Global Hierarchical Analysis (LGHA) attempts to discover accurate and relevant correlations from observational data (Lin and Orgun 2000, Lin and Orgun 2004, Lin et al 2000, Zhang et al 2006). There are three steps in LGHA. The first step involves a structural approach to find qualitative patterns from the given variables. Then, the second step applies a quantitative-based algorithm to find quantitative patterns from those variables. The third step generates global hybrid patterns by combining the local patterns obtained from the first two steps based on a certain criterion. LGHA enhances unsupervised learning by making the error of knowledge exploration as small as possible, especially when we are dealing with time series data or irregular data. LGHA also benefits from a visualization interface for the knowledg e explorers that could incorporate domain knowledge in the process.

In our framework, K-Means is applied to the dataset prior to the application of LGHA. K-Means efficiently clusters data so that valuable variables can be generalized into a worthwhile observational dataset. This will help improve the performance of LGHA which may otherwise be wasting a lot of time seeking latent variables and structural patterns constructed from the huge amount of original data.

The rest of the paper is organized as follows. Section 2 is a general review of unsupervised learning. Then we address the theories of K-Means and LGHA respectively, in sections 3 and 4. In section 5, we apply the two models in experiments with real-world datasets, Australian Medicare Data records. Also, the combination of K-Means and LGHA is shown to be feasible. Last section discusses related work and concludes the paper with a brief summary.

## 2. Unsupervised Learning

Traditional supervised learning attempts to define a

functional mapping between the given input and output. In a supervised system, the corresponding output vectors or patterns are reserved when the learner is provided with input patterns or vectors. In contrast to supervised learning, unsupervised learning presents a collection of unsupervised observations. The learner can neither recognize general regularity or explanation for these observations, nor generate any expected output. In fact, the purpose of unsupervisedlearning is to extract some understanding from these observations first, and then attempt to discover the latent variables.

Due to the fact that no supervised targets are expected in the stage of producing the output, finding new patterns from the original data is the main task of an unsupervised learner. Its distinguished learning and grouping features can avoid the case that useful information is treated as unstructured noise signal. In addition, new patterns provide more convenience of further analysis, i.e., in the subsequent knowledge exploration.

According to the literature (Engelbrecht 2002, Shawe-Taylor and Cristianini 2004, Shanahan 2000, Alpaydin 2004), most unsupervised learning approaches utilize statistical models. For example, a collection of original data *x1, x2…,xn* is received as input to an unsupervised system. We observe a data set *D={ x1,…,xN }* firstly. Then, we assume that there exists a model set $\Omega$, where $\Omega$ *={ 1, …, M }* (learner normally starts processing with some prior models *m*, where $m \in \Omega$; thus, $\sum_{m=1}^{M} P(m) = 1$ (Ghahramani =*1 m* 2004).

Given that all the models are based on the probability distribution over data points, that is:

$$P(m|D) = \frac{P(m)P(D|m)}{P(D)} \qquad (1)$$

Hence, the relationship between prior models and posterior models is given as:

$$P(m|D) = \frac{P(m)P(D|m)}{P(D)} \propto P(m) \prod_{n=1}^{N} P(x_n|m) \qquad (2)$$

Based on the theory of parametric estimation, all models can be formed by parametric probability distribution, i.e., each model *m* has its corresponding unknown variable parameter $\theta$. The best estimated value of $\theta$ over model *m, P( | m)* will be defined through the perspective of unsupervised learning.

Moreover, the satisfied parameter $\theta$ is obtained via the following equation (Vapnik 2000):

$$P(x|m) = P(x|q, m)\, P(q|m)dq \qquad (3)$$

Then, we map the model from the single data into the entire data set,

$$P(x\, D,|m) = P(x|\theta)\, P(\theta|D, m)d\theta\,, \qquad (4)$$

It is an arduous task to define a unique model being appropriate for all kinds of data, especially when the collection of observations is of high dimensionality. However, the dramatic increase of electronic information actually creates lots of high dimensional data. Thus, unsupervised learning offers two general approaches to deal with these high dimensional data, that is, dimensionality reduction and clustering.

In the next two sections, we discuss two models, the well-known K-Means Clustering Analysis (K-Means) and Local-Global Hierarchical Analysis (LGHA) (Lin and Orgun 2000, Lin and Orgun 2004, Lin et al 2000, Zhang et al 2006) that will be employed in unsupervised learning.

## 3. K-Means Clustering Analysis

### 3.1 Clustering Analysis

The basic idea of Clustering Analysis (CA) is to classify observations around different clusters in terms of their similarity measurement or distance measurement. Both of those measurements are implemented to compare relevancy between every two points of data, which are randomly chosen from a given data set. Thus, CA is suitable to accomplish two particular tasks. It on the one hand discovers homogeneousness among observations through calculating their internal cohesion. On the other hand, external separation of observations is also recognized

The model of CA can be divided into two distinct types based on the criterion of the choice of the number of clusters. Uncertain Cluster Analysis (UCA) performs exhaustive search to make clusters as precise as possible. It does not impose any limitations on the quantity, position or size of clusters. UCA describes clusters in more details than any other types. However, it may result in worthless or noisy clusters especially when the data set is huge and dynamic. Certain Cluster Analysis (CCA) achieves to group data by a predetermined condition of clusters. For example, given that a collection of data set *D={d1…dn}* and cluster number *k* or cluster size *q,* we follow the above chosen criterion to group the given data into *k* clusters or design clusters of size *q*.

CCA is considered to be an iterative algorithm in which the chosen data is moved among various cluster sets until the desired set is obtained (Jain and Dubes 1988, Han and Kamber 2001, Dunham 2003). Furthermore, the convergence of clusters directly influences the variety of the cluster's number. Less important data outside of known clusters are regarded as noisy or irrelevant data that can be removed.

Both of UCA and CCA contain three main steps to finalize the modelling of CA.

*Step One: Data Normalization*

Due to the difference in value, vector or dimension, the original data set is always considered as dirty data or unorganized data.

Thus, the importance of data normalization is significant. Three methods are commonly employed in this step, which are based on Sum, Standard Deviation and Maximum approaches.

To explain these three approaches, we assume a data matrix that contains *m* rows and *m* columns given below:

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{m1} & x_{m2} & \cdots & x_{mj} & \cdots & x_{mn} \end{matrix} \qquad (5)$$

√Sum Approach.

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} (i = 1,2, \ldots m;\ j = 1, 2, \ldots m \qquad (6)$$

where $\sum_{i=1}^{m} x'_{ij} = 1 \ (i,j = 1, \ldots m)$.

✓ Standard Deviation Approach

$$x'_{ij} \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i=1,2,\ldots m; j=1,2,\ldots m) \qquad (7)$$

where $\bar{x}_j = \sum_{i=1}^{m} x'_{ij}$ and $s_{ij} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x'_{ij} - \bar{x}_j)^2}$

✓ Maximum Approach

$$x'_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \quad (i=1,2,\ldots m; j=1,2,\ldots m) \qquad (8)$$

*Step Two: Distance Calculation*

Distance measure is a metric or quasi-metric on the data space used to quantify the similarity of patterns (Jain et al 1999). Hence, we calculate all possible distances between any two given data points *xi and xj* from the given data set *X = {x1 … xn }, i, j ∈n* in this step.

Commonly, two approaches, Euclidean distance and Mahalanobis distance are introduced to perform this task.

✓ Euclidean distance

It is concerned with the distance between two data points, *xis* and *xjs.* And the distance will be defined as the sum of the component-wise differences squared.

$$dij = d(x_i,x_j) = (x_{i1},x_{j1})^2 = (x_{i2},x_{j2})^2 + \ldots + (x_{is},x_{js})^2 \qquad (9)$$

✓ Mahalanobis distance

It is a measure of distance between two random feature subset points *xi* and *xj*.

$$dij = \sqrt{(m_i,m_j)^T \Sigma^{-1} (m_{i1},m_{j1})} \qquad (10)$$

where *mi=E(xi)*, *mj=E(xj)*. $\Sigma i = E\{(xi-mi)(xi-mi)^T\}$ and $\Sigma j = E\{(xj-mj)(xj-mj)^T\}$ are the covariance matrices. Thus, $\Sigma = p\Sigma i + (1-p)\Sigma j$, where *p* is the Minkowsky's coefficient (it is generally nominated as the value of 1 or 2).

*Step Three: Choose an Algorithm*

A clustering algorithm is understood as a computational procedure such that, given a data set, it organizes similar data inside the range closer to the cluster point and dissimilar outside. K-Means is one of the most efficient and well-known algorithms for finding and constructing the cluster structure according to a formal definition or function. The following section discusses the way in which K-Means Clustering Analysis efficiently works for the CA model.

### 3.2 K-Means Clustering Analysis

The essential idea behind clustering analysis is to organize similar data into various groups (clusters) according to certain criteria. As a result, the distance between each pair of data points within a cluster has less distance than those between a pair of data points belonging to various clusters or any data outside of the cluster.

K-Means Clustering Analysis (which we will refer to as K-Means) is a clustering model that produces a partition of the data set into non-overlapping clusters along with within-cluster centroids (Mirkin 2005). It is an easy, fast and memory-efficient method to compute from a given data set a certain fixed number of clusters *K*. And the means are the aggregate representations of clusters and as such they are sometimes referred to as standard centroids. Hence, K-Means is to define *k* centroids distributed to each cluster.

For example, in a given data set $D=\{X_1,\ldots,X_n\}$, an integer value *k* and a set of centroids represented as $C=\{C_1 \ldots C_k\}$, the main task of K-Means is to define a mapping $d:D \rightarrow \{1,\ldots,k\}$ where each *xi* is assigned to one cluster $Kj, \leq 1 \leq j \leq k$, where $K_j=\{ t_i \mid d(x_i) = C_k$, for $1 \leq I \leq n, xi \in D, ck \in C\}$.

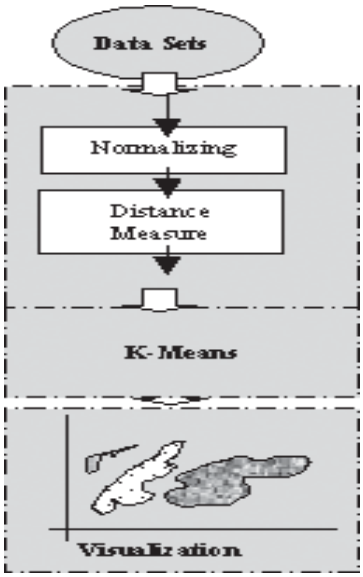


Figure 1. The Framework of K-Means Clustering Analysis

We initially locate *k* centroids and then assign data into the closest centroid. We relocate the positions of *k* centroids to make them more precise while all data are assigned around the centroids. K-Means algorithm halts when no more centroids need further adjustment:

$$k = \sum_{i=1}^{n} \sum_{j=1}^{k} d(x_i, c_j) \qquad (11)$$

Here *n* is the scope of data, *k* is the number of centroids and $d(x_i, c_j)$ is the function to learn the distance between a data point and its closest centroid. As a result, K-means needs to timely upgrade both in the position of centroids and the distance between data points and the centroids. As K-means is a well-known algorithm, we omit further details.

## 4. Local-Global Hierarchical Analysis

### 4.1. Hierarchical Analysis

Hierarchical analysis (HA) is an advanced statistical model based on the polynomial approach (Lin and Orgun 2000; Lin and Orgun 2004). It is employed to investigate the latent relationships between an individual data point and the entire data set. Thus, a lot of computation and validation at the inter-level and intra-level are required when the numbers of variables at different levels are enormous.

In HA, the original data set is first restructured into the hierarchical variable set. We first assume a dependent variable. Subsequently, the rest of the variables are measured as independent variables. The first level (the entry level or the lowest level) measures the dependent variable while other independent variables are estimated by the following levels.

The HA model usually commences via a pure parameter model (also called null variable model), and only an intercept parameter is assumed in this initial model,

$$Y_{ij} = \gamma 00 + u_{oj} + e_{ij} \qquad (12)$$

where, $\gamma 00$ is the models intercept, $u_{oj}$ is the random higher level effect with variance $\sigma^2 u0$ and $e_{ij}$ is the first level effect

with variance $\sigma^2 e$. The subscripts *i* and *j* stand for variables in the first level and the higher level, respectively.

Then we import each independent variable *Xp* (*p* is the number of independent variable *X*) into the above base model,

$$Y_{ij} = \gamma 00 + \gamma_{p0} X_{pij} + u_{0j} + e_{ij} \quad (13)$$

Due to the fact that the dependent variable is unchangeable, the corresponding variance components of slopes are established at zero value. We examine the difference of deviance between this model and the pure parameter model so that we can assess whether the improvement is enough to yield an optimal model.

Now we can add the higher level independent variables into the model, as in equation

$$Y_{ij} = \gamma 00 + \gamma_{p0} X_{pij+} \; {}_{oq} Z_{qj} + u_{0j} + e_{ij} \quad (14)$$

where *Z* is the other independent variable in the higher level, and the subscript *q* is the number of *Z*. We can use one of parameter estimation methods to relocate the parameters in order to fit the complicated model optimally. In addition, one more step of estimation (so called "cross-level interaction") is required to fit the following equation if the number of *p* and *q* is huge.

$$Y_{ij} = \gamma 00 + \gamma_{p0} X_{pij} \gamma_{oq} Z_{qj+} \gamma_{pq} Z_{qj} X_{pij} + u_{0j} + e_{ij} \quad (15)$$

#### 4.2. Local-Global Hierarchical Analysis

The model of Local-Global Hierarchical Analysis (LGHA) is built by three levels, namely, Local Structure-based Level (LSL), Local Value-based Level (LVL) and Global Pattern Level (GPL) (Lin and Orgun 2000, Lin and Orgun 2004, Zhang et al 2006). In the first level, LSL, distance measures are adopted by a structural search on the data set through polynomial modelling in order to group qualitative patterns. In the second level, LVL, statistical measures are employed to extract conditional distributions from the data set to define quantitative patterns. In the last level, GPL, patterns from the previous two local levels are combined into global patterns to accomplish unsupervised knowledge exploration of latent relationships.

n terms of the theory of HA, we define data in three parts, namely the quantitative part, the qualitative part and the position part. Hence, every data set *D* has a triple value defined for it:

$$D = \{v, k, p\} \quad (16)$$

Here *v* is the quantitative part, *k* is the qualitative part, and *p* is the position part.

Then two assorted patterns are created to complete local analysis. One is the quantitative pattern that corresponds to every data both in *v* and *p*

$$V = \{v, p\} \quad (17)$$

And the second is the qualitative pattern that corresponds to every subset both in *k* and *p*.

$$Q = \{k, p\} \quad (18)$$

Under the environment of LGHA, we assume that there exists a finite data set $D=\{d_1 \ldots d_m\}$, where $t_i$ is their position parameter. So we can transfer the original data into a new observational group.

$$D = (t_i) = \{di, ti\} \; (1 \leq \; \leq n) \quad (19)$$

where *D(ti)* denotes the value of the corresponding data item by its position parameter. For example, we randomly choose 1000 data items from the original set. In this example, *m* is 1000 and *ti* ranges from $t_1$ to $t_{1000}$ and is matched with each chosen data item. Each new observational object is expressed as $D(t_i)=\{d_i,t_i\}$ $(1 \leq i \leq 1000)$.

Now we combine every two consecutive objects into another new observational group $S(t_i)=\{D(t_i),D(t_{i+1})\}$. There exist three states in the new group depending on whether values increase, decrease or keep the same. We suppose that $S_s$ is the same value as the prior one; $S_u$ is the stronger value compared with the prior one (the value has increased); and $S_d$ is the weaker value compared with the prior one (the value has decreased). Hence, the new group can be expressed as

$$S(t_i)=\{D(t_i),D(t_{i-1})\} \; (1 \leq i \leq n) \quad (20)$$

$$S(t_i)=\{S(t_{di}) \; ti\} \; (1 \leq i \leq m) \; and \; S(t_{di}) \in \{S_s, S_u, S_d\} \quad (21)$$

As it has been mentioned above, given a finite amount of data $D = \{d_1, d_2 \ldots d_m\}$, we can divide it into three parts: the quantitative part, the qualitative part and the position part. Then, the qualitative patterns and the quantitative patterns are separately created by those three parts. The qualitative one is a base distribution in probability space. And the quantitative one is a coefficient distribution of the data.

Therefore, the task of unsupervised learning can be formulated as follows:

$$D = \{V\} \otimes \; ^{\Gamma}Q\} \quad (22)$$

The quantitative pattern of the data set is

$$V = \begin{Bmatrix} D(t_1) & D(t_2) & \cdots & D(t_{i-1}) \\ D(t_2) & D(t_3) & & D(t_i) \end{Bmatrix} \quad (23)$$

And the qualitative pattern of data set is

$$Q = \{ S(t1), S(t2), \ldots S(ti)\} \quad (24)$$

Therefore, data set under LGHA can be written in the form shown below

$$D_m = \{D(t_1) \otimes S(t_1), D(t_2) \otimes S(t_2) \ldots, D(t_{m-1}) \otimes S(t_{m-1})\} \quad (25)$$

Also we assume that for each successive pair of data, it has a uniformly distributed function

$$f_c = t_{i+1} - t_i \quad (26)$$

The main advantage of LGHA is to effectively increase the accuracy of unsupervised learning because the final result of knowledge exploration integrates the completed analysis from three levels.

## 5. Hierarchical Analysis in Unsupervised Learning

In this section, Local-Global Hierarchical Analysis (LGHA) is utilized to assist unsupervised learning. In recent years, the discovery of two types of data, named complete/partial similarity data and periodicity data has become the focus of attention, especially when there are high-dimensional features or time-varying attributes in the data.

According to LGHA, we consider observed data as expressed in two patterns. The first one is the qualitative pattern, also called structural pattern, whose focus is on analysis of state space $S = \{S_s , S_u , S_d\}$. The second part is the quantitative pattern, also called pure-value pattern, whose focus is on the analysis of probability space *P*. Eventually, we combine the results from the above two levels into a hybrid pattern to obtain the final learning result.

### 5.1. Qualitative Patterns

The generation of the qualitative pattern is based on state *S* processed in data. We first suppose that a qualitative sequence on *Q* is a set of structural vector sequences, $Q=\{Q_1,\ldots,Q_m\}$, where each $Q_{ti}=(S(t_1), S(t_2),\ldots, S(t_{tn}))$ $(1 \le i \le m)$ denotes the *m*-dimensional attributes for each *Qi* that is to be assigned to a specified distribution cluster.

Hence, we can consider $\{Qt_i: i \in m\}$ as a qualitative pattern.

$$Q_{ti} = \{(S(t_{f1}),t_1), \ldots (S(t_{fm}),t_m)\} \quad 1 \le i \le m \qquad (27)$$

$$S(t_{f1}) = \in \{S_s, S_u, S_d\} \qquad (28)$$

where *Qt* is an irreducible homogeneous qualitative set with states of *S*.

Now we can define a probability set which is a correlated measure of the relationship between two data sequences. It is called a correlation ratio sequence for all qualitative states and given as follows.

$$w_{jk} = P\left(\frac{(Q_t(S(t_{fi}))}{(Q_t(S(t_{fi}))}\right). \qquad (29)$$

Hence, we can find a unique, strictly positive statistical distribution for each *Qt.*

### 5.2 Quantitative Patterns

We assume $\{Y(t_i): i \in m\}$ is the quantitative pattern, where

$$Y_{ti} = \{d_i, h(d_i)\} \qquad (30)$$

Due to quantitative-based search, the unknown regression function *h(di)* is obtained by applying a Taylor expansion of order *p* in a neighborhood of *di-1* with its remainder *np*.

$$h(di) = \sum_{m=0}^{p} \frac{h^{(m)}(d_{i-1})}{m!}(d_i-d_i\text{-}1)^m + n_p = \sum_{m=0}^{p} \beta m(d_i-d_i\text{-}1)^m + n_p \qquad (31)$$

where $$n_p = w_p \otimes d_p \qquad (32)$$

Also, least square estimation can make the value of $\beta$ under the linear model

$$\hat{\beta} \sim p(\beta, \sigma^2 a_{mm}). \qquad (33)$$

where *amm* is the *m*th diagonal subsets in matrix *N*. In short, we can transform the quantitative pattern problem into a local linear model and formulate it as the data distribution functional analysis.

We observe that quantitative pattern search has two benefits. It helps remove non-relevant data by a near zero local linear model and groups similar ones into various valuable clusters. It also sheds light on whether observational data is linearly separable; it will still be linearly separable even when some data are marked as redundant and deleted out of observational groups.

### 5.3 Hybrid Patterns

We combine the above two kinds of patterns to discover hybrid patterns from a given data set. Firstly, we extract the qualitative pattern $\{Q_{ti}: i \in m\}$ to apply the data functional distribution sequence on the state space *S*. Secondly, we suppose the quantitative pattern $\{Y_{ti}: i \in m\}$ is a nonnegative random vector process. Thirdly, we apply the conditional distribution of feature $F_t$. For example, if $Q_{ti} = S(t_{fi})$, $Y_{ti}$ has a Poisson distribution with mean $\lambda i$, let $E(Yt_i/Q_{ti})$, the conditional mean of $Y_t$ can be calculated by the following function:

$$u(t) \sum_{i=1}^{m} \lambda i \, D(t_i) \qquad (34)$$

where the random data $D(t_i)$ is the indicator of the event $\{Q_{ti} = S(t_{fi})\}$. At the same time, the state development probabilities are then given as shown below:

$$\pi i = \frac{e^{-\lambda i}\lambda i}{f_i!} \qquad (35)$$

Now, according to the above deduction, the hybrid pattern is defined as Poisson Hidden Models and the hybrid pattern search becomes the problem of Poisson distribution.

## 6. Empirical Study

### 6.1. Background

This section presents an experimental study on real-world datasets to test two analytical models, K-Means Clustering Analysis (KMeans) and Local-Global Hierarchical Analysis (LGHA). In our experiments, the chosen dataset is the diabetic segment data from Australian Medicare Database that has been collected in Australia domain-wide since the inception of Medicare in 1975. Diabetes occurs when the glucose enters bloodstream from food or drink, and cannot be processed by human's body itself. It causes a build-up of sugar in the blood. Without treatment, it can cause many other side effects. People with diabetes can neither produce enough insulin to meet their requirements nor can their cells respond properly to the insulin. As a result, the glucose builds up to abnormal levels in their blood.

The diabetes data in Medicare transaction during period 1997 and 1998 will be utilized for our knowledge exploration through unsupervised learning. The primary goal of this application is to generally discover certain implied diabetes patterns. Two tasks will be attempted. The first one is to distinguish patterns hidden in longitudinal records of those diabetes patients. The second one is to explore some interrelated patterns of care between patients and doctors.

The data used in this case study is extracted from the Medicare transactional database that consists of records of 10,000 diabetic patients using Medicare services paid by

| Encrypted Provide Number | Encrypted PIN | Method of Payment | Class Number |
|---|---|---|---|
| ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ |
| Date of Service | Benefit | Reason Code of Rejection | Referral Provider |
| ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ |
| Processing Indicator | Date of referral | Hospital | ……. |
| ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | ~~~~~~~~~~~~~~~ | (…….) |

Table 1. The Medical Record of Diabetic Patents

H.I.C. under the Medicare Benefits Schedule. The data extracted from Medicare is raw transaction data, which is a very large data set with millions of records and each record has more than a hundred feature subsets.

For computing and storage purposes, the identification is set by a particular item number of every service record in the raw database. As a result, those kinds of item numbers are information carriers of the patients' medical service during years 1997 to 1998. The fields in that medical information includes Encrypted Provider Number, Encrypted PIN, Date of Service, Benefit, Processing Indicator, Date of referral, Method of Payment, Class Number, Reason Code of Rejection, Referral Provider, Hospital. We list them in Table I.

### 6.2. Empirical Study of K-Means Clustering Algorithm

As we mentioned above, K-Means assigns data into *k* clusters according to a distance criterion. It implements the distance comparison to every pair of data points to find their relevance and distinguishes them between inter-cluster and intra-cluster. Therefore, the results from K-Means are the embodiment of cohesion of homogeneousness of data and separation of disparate data.

In this experiment, six categories of data are chosen as experimental data, and are represented by *X1* (Various Providers), *X2* (Methods of Payments), *X3* (Type of Items), *X4* (Days of Treatments), *X5* (Benefits) and *X6* (Various Referral Providers). The data shown in Table II is some sample records of patients.

Distance Matrix =

$$\begin{pmatrix} X1\ X1 & X2\ X1 & X3\ X1 & X4\ X1 & X5\ X1 & X6\ X1 \\ X1\ X2 & X2\ X2 & X3\ X2 & X4\ X2 & X5\ X2 & X6\ X2 \\ X1\ X3 & X2\ X3 & X3\ X3 & X4\ X3 & X5\ X3 & X6\ X3 \\ X1\ X4 & X2\ X4 & X3\ X4 & X4\ X4 & X5\ X4 & X6\ X3 \\ X1\ X5 & X2\ X5 & X3\ X5 & X4\ X5 & X5\ X5 & X6\ X4 \\ X1\ X6 & X2\ X6 & X3\ X6 & X4\ X6 & X5\ X6 & X6\ X6 \end{pmatrix}$$

| **Patient's Number** | **Various Providers** | **Methods of Payment** | **Type of Items** | **Days of Treatments** | **Benefits** | **Various Referral Providers** |
|---|---|---|---|---|---|---|
| **1** | **3** | **1** | **17** | **40** | **1197.1** | **1** |
| **2** | **3** | **1** | **9** | **7** | **220.5** | **1** |
| **3** | **6** | **2** | **12** | **5** | **345.75** | **4** |
| **4** | **3** | **1** | **6** | **7** | **202.55** | **1** |
| **5** | **4** | **3** | **7** | **9** | **375.85** | **1** |
| **6** | **6** | **2** | **8** | **11** | **410.1** | **2** |
| ……. | ……. | ……. | ……. | ……. | ……. | ……. |

Table 2. Sample records of Diabetic Patients

| X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|
| -0.0736 | -0.0196 | 0.0736 | 0.4576 | 0.06212244 | -0.024 |
| -0.0736 | -0.0196 | -0.0864 | -0.2024 | -0.13319756 | -0.024 |
| -0.0136 | 0.0004 | -0.0264 | -0.2424 | -0.10814756 | 0.036 |
| -0.0736 | -0.0196 | -0.1464 | -0.2024 | -0.13678756 | -0.024 |

Table 3. Normalized Records

| General Patient | Professional Attendances | Diagnostic Services | Approved Dental | Diagnostic Imaging | Pathology Services |
|---|---|---|---|---|---|
| GP | PA | DS | AD | DI | PS |
| Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |

Table 4. Six Categoreis of Benefits Schedule

| Patient Name | Patient Number | Date of Service | Record from Schedule | Item Number |
|---|---|---|---|---|
| ~~~~~~~ | ******** | ######## | PA | 2 |
| ~~~~~~~ | ******** | ######## | GP | 1 |
| ~~~~~~~ | ******** | ######## | DI | 5 |
| ~~~~~~~ | ******** | ######## | DS | 3 |
| ~~~~~~~ | ******** | ######## | PS | 6 |

Table 5. Medical Records of Sample Patients

In the beginning, the original record set of patients is normalized as shown in Table III. After normalizing the observational data, we calculate all possible distances between two arbitrary data points and quantify similarity patterns. The values of all distances are placed into a matrix.

We transform the distance matrix into observational matrix, as shown below.

$$\text{Observational Matrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 \\ 19 & 20 & 21 & 22 & 23 & 24 \\ 25 & 26 & 27 & 28 & 29 & 30 \\ 31 & 32 & 33 & 34 & 35 & 36 \end{pmatrix}$$

Then the K-Means algorithm acquires a predetermined value of *k* and starts to compute similar patterns in distance matrix (details are not shown). The results (shown in Table VI) after clustering are sent to the visualization stage so that we can get an intuitive understanding of the results with the aid of the visual user interface. (See Fig. 2)

Now some particular characteristics are discovered from the application of the K-Means algorithm.

We find that sub-matrix *16(X4X3), 17(X5X3), 21(X3X4 )* and *23(X4X5 )* appear in cluster two. It can be understood that some moderate interrelationships exist among variables *X3, X4* and *X5*. In other words, types of items, days of treatment and benefits can be influenced by each other. Furthermore, there appears to be a stronger correlation among those three variables while *X1* is considered as an indexed observation.

Due to the fact that *X1* has obvious relevance with *X3*, *X4* and *X5*, we may indicate the tendency of varying of *X3* (type of items), *X4* (days of treatment) and *X5* (benefits) through tracking changes to *X1* (various providers), and vice versa. For instance, we could make a predictive statement to the patients that they probably need to pre-arrange their scheme of treatment with different medical providers based on their historical records of types of items, days of treatment or benefits.

Variable *X2* shows a similar linear relationship with *X1, X5* and *X6*. In other words, the method of payment is altered at the same time with the changes of various providers, benefits and various referral providers. Thus, we only need to trace the distribution of methods of payment while we attempt to indicate the regularity of the other three variables. The figure 2 provides intuitive evidence to the idea. There *X1* and *X6* show a stronger linear correlation with *X2*; it might therefore imply that the medical provider (*X1* Various Provider and *X6* Various Referral Provider) is the most important factor to influence the patients' choice of their payment methods. In addition, this correlation could be partly used in the predictive model that indicates the financial status of patients.

### 6.3. Empirical Study of Local-Global Hierarchical Analysis

According to the knowledge we acquired from K-Means, some specific variables are selected from the observational data set.

| | | | | | | |
|---|---|---|---|---|---|---|
| Group around Cluster One | | | | | | |
| Sub-matrix | 3 | 4 | 5 | 13 | 6 | 19 |
| | 27 | 28 | 30 | 31 | 35 | |
| Group around Cluster Two | | | | | | |
| Sub-matrix | 9 | 10 | 12 | 14 | 20 | 32 |
| Group around Cluster Three | | | | | | |
| Sub-matrix | 1 | 17 | 8 | 15 | 16 | 18 |
| | 21 | 22 | 24 | 25 | 29 | 33 |
| | 34 | 36 | 23 | | | |
| Ungrouped | | | | | | |
| Sub-matrix | 2 | 7 | 11 | 16 | | |

Table 6. The Observations Grouped into Three Clusters

These data show the patients' status of treatment, such as the times of visiting consultants and so on. Those irregular attributions in data make analysis difficult. Hence, we define the date of a medical service as an index of hierarchical patterns, and choose six general categories of medical services. They are Professional Attendances (PA); Diagnostic Services (DS); Approved Dental Practitioner Services (AD); Diagnostic Imaging Services (DI); and Pathology Services (PS). As shown in Table IV.

Also, we sample patient's records and process them in LGHA. Table V (given on the previous page) shows some sample patient's records.

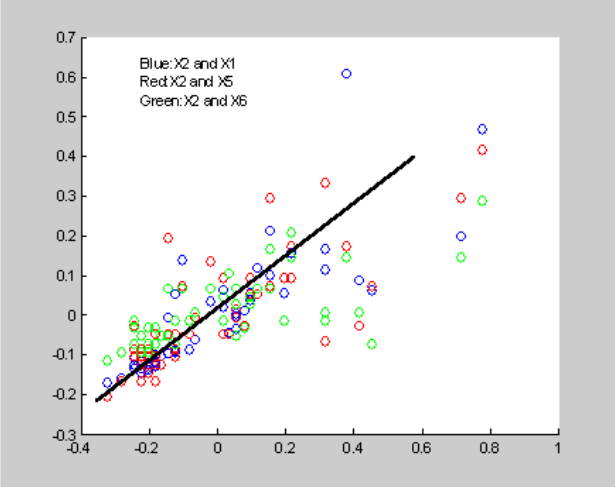


Figure 2. The Relationship among *X2* and *X1*, *X2* and *X5*, *X2* and *X6*

According to the theory of LGHA, we firstly investigate the qualitative pattern on state-space *Si={Su, Ss, Sd}* because only three states are designated for our analysis.

The algorithm of qualitative pattern search is shown in Table VII.

| | | |
|---|---|---|
| ***Input:*** | ***F*** | ***% Original data (f1…fm)*** |
| | ***S*** | ***% Sequence set of states*** |
| | ***t*** | ***% Position parameter*** |
| | ***r*** | ***% Ratio set of state transition*** |
| | ***w*** | ***% Probability set of state ratio*** |

***Procedure Qualitative Pattern Search***

*Begin*
*Define Q{S;t}*
*Define ω {r;t}*
*Initial $S_0$, Initial t*
*For Int i=2;i<=m;i++*
*S[t].t=i-1*
*If fi=fi-1 then S=S + Ss;*
*Else if fi<fi-1 then S= S + Sd*
*Else if fi>fi-1 then S= S + Su*
*Initial r, Initial t*
*For Int i=2;i<=m;i++*
*ω [t].t=i-1*
*r=r+(Q{S[i]}/Q{S[i-1]})*
*[t]=+P(r) % the value of transition probability*
*Plot [t]*
*End*

***Output relationship among states***

Table 7. The Algorithm of Qualitative Pattern Search

Since each patient record length is different in those health records, we can only use their statistical value as variables in the linear regression model. The algorithm of quantitative pattern search is shown in Table VIII.

| | | |
|---|---|---|
| ***Input:*** | ***F*** | ***% Original data (f1…fm)*** |
| | ***Y*** | ***% Quantitative Pattern*** |
| | ***Q*** | ***% Set of frequency distribution*** |
| | ***H*** | ***% Set of quantitative data*** |
| | | ***% An auto-regression model*** |

***Procedure Quantitative Pattern Search***

*Initial the value of Q {fi, h(fi)}, k= numbers of clusters*
*Sort the item numbers into quantitative pattern*

$$Compute\ h\ (f_i) = \sum_{m=0}^{p} \frac{h^{(m)}(f_{i-1})}{m!} (f_i - f_i - 1)m + n_p = \sum_{m=0}^{p} \beta\ m(f_i - f_{i-1})^m + n_p$$

*Let Yt(i-1)=Q(fi)-Q(fi-1),*
*Let Yti=Yt(i-1)+*
*Compare Yti and Yt(i-1)*
*If $Y_{ti}$ is distinct then send to H*
*Plot H*

***Output relationship among quantitative pattern***

Table 8. The Algorhithm of Quantitative Pattern Search

Lastly, we combine the two patterns from the qualitative and quantitative parts into a hybrid model. Then some results of our experiments can be explained as follows:

Items 4 and 5 are independent. We have found that there exist some similarities of distribution between them.

However, non-related elements exist between their clusters. This means that the patients have received a number of treatments that are similar but in different time periods.

Items 1 and 6 have moderate similarity. This means that doctors probably have different levels of knowledge of diabetes problems.

## 7. Conclusion and the future work

In recent years, various studies have been conducted in knowledge exploration from massive real-world datasets for searching different kinds of and/or different levels of patterns. However, the techniques and methods developed in those studies are not for general cases. For example, most researchers use statistical techniques such as, Model-based technique, or a combination of several techniques to search different pattern problems such as in periodic pattern searching, and in similarity pattern searching.

Agrawal and others (Agrawal et al 1995) present a "shape definition language", called *SDL*, for retrieving objects based on shapes contained in the histories associated with these objects. Das with others (Das et al 1997, 1998) describe adaptive methods, which are based on similar methods for finding rules and discovering local patterns. Williams and others (Huang 1997, Williams et al 2001) have considered three alternative feature vectors for representing variable-length patient health records.

Our work is different from theirs. We have paid more attention to knowledge exploration through unsupervised learning. Unlike a supervised learning system that receives predetermined input and produces desired output, an

unsupervised learning system emphasizes the process of understanding current data and mining latent variables/ relationships from the learner system. Due to the fact that no specific results are expected, an un unsupervised learner can therefore deal with large scale data, high dimensional data and dynamic data.

In this paper, we address two effective models, K-Means Clustering Analysis (K-Means) and Local-Global Hierarchical Analysis (LGHA) and combine them into a single framework to assist unsupervised learning. Their comprehensive algorithms effectively implement dimension reduction, grouping observations and exploration of hidden knowledge. We designed three steps to build a complete K-Means model. In addition, we employ this K-Means model into the real-world experiment. The result is satisfying because it extracted some valuable underlying relationships by partitioning the unorganized data into predetermined clusters. The clustering patterns are also visualized for more intuitive explanations.

We also build on a comprehensive structure and modeling procedure of LGHA (Lin and Orgun 2000, Lin and Orgun 2004). Three levels are defined to build the LGHA model. The transition probability is utilized in the first level to define a conditional distribution. The quantitative patterns are placed in the second level to extract pure value point data. In the final level, we combine qualitative and quantitative patterns to obtain a global hybrid pattern. Particularly, the empirical study suggests that LGHA has stronger ability in exploring quantitative correlations between observational variables.

We have examined the prospect of a combined system that can employ the model of K-Means and LGHA together. The results from our empirical studies are encouraging. In summary, structural learning (the recognition of latent variables) within LGHA might be efficiently enhanced by powerful clustering ability of the K-Means model.

In future work, we plan to apply our unsupervised learning approach to other real-world data such as financial datasets and taxation datasets. Furthermore, the combined system of K-Means and LGHA will be studied in more detail.

**Acknowledgements**

The work reported in this paper is partially supported by the Australian Research Council under the Linkage project scheme (LP0561985) and the Capital Market CRC Limited (CMCRC). Thanks are also due to Warwick Graco for stimulating discussions.

**References**

[1] Agrawal, R., Psaila, Wimmers, G. E. L., Zait, M (1995). Querying shapes of histories. *In*: Proceedings of the 21st VLDB Conference.

[2] Alpaydin, E (2004). Introduction to machine learning. MIT Press.

[3] Das, G., Gunopulos, D., Mannila, H (1997). Finding similar time seies. *In*: Principles of Knowledge Discovery and Data Mining '97, 1997.

[4] Das, G., Lin, K., Mannila, H., Renganathan, G. and Smyth, P (1998). Rule discovery from time series. *In: Proceedings of the international conference on KDD and Data Mining (KDD-98).*

[5] Dunham, M. H (2003). *Data* Mining Introductory And Advanced Topics. Pearson Education Inc., Upper Saddle River, New Jersey.

[6] Dy, J.G., Brodley, C.E (2000). Feature Subset Selection and Order Identification for Unsupervised Learning, *Proc. 17th Int'l Conf. Machine Learning*, p. 247-254.

[7] Dy, J. G., Brodley, C. E. (2004) Feature Selection for Unsupervised Learning, *Journal of Machine Learning Research 5*. 845-889.

[9] Engelbrecht, A. P (2002). Computational Intelligence: An Introduction. J. Wiley & Sons, Chichester, England.

[10] Ghahramani, Z (2004). *Unsupervised learning.* Advanced lectures on Machine learning, Lecture Notes in Artificial Intelligence 3176. Springer-Verlag.

[11] Han, J., Kamber, M (2001). Data Mining: Concepts and Techniques. Morgan Kaufman.

[12] Huang, Z (1997). Clustering large data set with mixed numeric and categorical values. *In:* 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[13] Jain, A. K., Murty, M. N., Flynn, P. J. (1999) *Data Clustering: A Review*. *ACM Computing Surveys*, 31 (3).

[14] Jain, A. K., Dubes, R. C. (1988). *A*lgorithm for Clustering Data, Chapter Clustering Methods and Algorithms. Prentice-Hall Advanced Reference Series.

[15] Jordan, M. I., Jacobs, R. A (1994). Hierarchical Mixtures of Experts and EM Algorithm, *Neural Computation*, 6. 181-214.

[16] Kohavi, R., John, G. H. (1997). Wrappers for Feature Subset Selection, *Artificial Intelligence*, 97 (1-2) 273-324.

[17] Lin, W. Q., Orgun, M. A (2000). Temporal data mining using hidden periodicity analysis. In *Foundations of Intelligent Systems, 12th International Symposium, ISMIS 2000,* Charlotte, NC, USA, October 11-14, 2000, Lecture Notes in Computer Science 1932, p. 49-58.

[18] Lin W. Q., Orgun M. A (2004). Structural pattern discovery in time series databases. *In:* S-H. Chen and P. P. Wang (editors), Computational Intelligence in Economics and Finance, Springer-Verlag, Berlin Heidelberg, p. 262– 287.

[19] Lin, W. Q., Orgun, M. A., Williams, G (2000). Temporal data mining using multilevel-ploynomial models. *In:* Intelligent Data Engineering and Automated Learning - *IDEAL 2000,* Shatin, N.T. Hong Kong, China, December 13-15, 2000, Lecture Notes in Computer Science 1983, p. 180-186.

[20] Liu, H., Motoda, H (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers.

[21] Liu, H., Yu, L (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17 (3) 1-12.

[22] Mirkin, B. G (2005). *Clustering for Data Mining : A Data Recovery Approach.* Chapman & Hall CRC.

[23] Shawe-Taylor, J., Cristianini, N (2004). *Kernel methods for pattern analysis*. Cambridge University Press, New York. Shanahan, J. G. (2000) *Soft computing for knowledge discovery: introducing Cartesian granule features.* Kluwer Academic.

[24] Vapnik, V. N. (2000) The nature of statistical learning theory. Springer.

[25] Williams, G., Baxter, R., He, H (2001). Feature selection for temporal health records. *In:* Knowledge Discovery and Data Mining - PAKDD 2001, 5th Pacific-Asia Conference, Hong Kong, China, April 16-18, 2001, Lecture Notes in Computer Science *2035,* p. 198-209.

[26] Zhang, Y., Orgun, M., Lin, W., Graco, W (2006). An Application of Time-Changing Feature Selection. *In* :Graham J. Williams.

[28] Simeon J. Simoff (Eds.): *Data Mining - Theory, Methodology, Techniques, and Applications.* Lecture Notes in Computer Science 3755, p. 203-217, Springer Berlin.