이학 석사 석사학위논문

Development of Prediction Models using Clustering

Clustering을 이용한 예측 모형 개발

2014년 2월

서울대학교 대학원

통계학과

민병주

Abstract

Development of Prediction Models using Clustering

Byung ju Min
The Department of Statistics
The Graduate School
Seoul National University

Repeated measures data have been commonly generated in many clinical studies. One of main objectives of the repeated measures data analysis is to predict a future outcome from the previously observed values. Repeated measures data often show certain patterns over time which can be easily checked by simple scatter plots. In this paper, we demonstrate that the use pattern information increases the accuracy of prediction in repeated measures data analysis. We propose to make a prediction model first by clustering data patterns via clustering methods and later by adding this clustering information into our model as a variable. We illustrate our approach using a real clinical data for bipolar patients. One of the clinical outcomes is Clinical Global Impression (CGI) value to predict patient's the extent of depression. We chose the better measure from calculated distances, Euclidean and 1-corr, between individual CGI values and clustered them by hierarchical clustering methods. Then we developed best prediction model of the extent of depression via above results. Here, we used linear mixed effects model in order to consider the effect of individual by adjustment of random effect. In terms of relative quality and prediction, our proposed method outperformed the models without clustering information according to AIC standards and prediction error.

…………………………………….

Keywords : clustering method, linear mixed effects model, prediction model

Student Number : 2012-20224

Contents

1. Introduction

Bipolar disorders (BD) are chronic conditions of an episodic and recurrent nature (Judd et al., 2002). Previous studies have reported 19 - 76% of patients with BP suffer from constant psychosocial impairment during the euthymic phase (Marangell et al, 2009; Tohen et al, 2000). Long-term treatments of episodes and early detection have been identified as key factors in the effective medical care of BP (Bauer et al, 2008; Swann, 2005, Moon et al, 2012). In addition, the probability of suicide increases if the diagnosis and prophylaxis are delayed; accurate diagnosis and prediction should be encompassed in the key factors. There are several tests to diagnose BD: ① Hamilton depression rating scale (Ham - D), ② Montgomery - Asberg depression rating scale (MADRS), ③ Young mania rating scale (YMRS), ④ bipolar depression rating scale (BDRS), ⑤ clinical global impression scale for bipolar disorder (CGI - BP) (Demyttenaere and Fruyt, 2003; Montgomery and Asberg, 1979; Hamilton, 1960; Young et al,1978; Berk et al, 2007; Spearing et al,1997). These measures provide clinical information on BD treatment status. Especially, the CGI - BP was developed for use in National Institute of Mental Health (NIMH) - sponsored clinical trials to provide a brief, stand-alone assessment of the clinician's view of the patient's psychosocial functioning prior to and after initiating a study medication. The CGI-BP provides a clinician-determined summary measure that takes total available information into account; this includes a knowledge of the patient's history, psychosocial circumstances, symptoms, behavior, and the impact of the symptoms on the patient's ability to function (Busner et al, 2007).

Repeated measures data have been generated and used in various fields. In real clinical studies, many variables representing patients' clinical status are repeatedly observed over different time points. For example, in our BP study, Patients update their clinical progress every month with new CGI-BP measures; while CGI values can be categorized to CGI-Severity (CGI-S) and CGI-Improvement (CGI-I), the former measures the current status of a patient while the latter measures the improvement between the patient's two CGI-S measures, our data consists of CGI-S values only. The main goal of generating repeated measures data especially over time is to predict the future outcome.

There have been a huge number of statistical models have been proposed for analyzing repeated measures data: multivariate regression model and mixed effects models for continuous outcomes and generalized estimating equations models for discrete outcomes (Cnaan et al, 1997).

When we analyze repeated measures data over time, there are some cases when the data show certain patterns. If we can effectively identify these patterns well, they can be quite informative in predicting future outcomes. However, if the patterns are mixed together, some patterns might not be well demonstrated in a scatter plot with entire data-points. Thus, it would be difficult to identify these patterns.

In this paper, we propose a simple prediction method to take advantage of the patterns in the data to make a better prediction of future outcome. The idea is very simple. Our proposed method first uses clustering algorithms to identify patterns in the data and later uses these patterns to make a prediction model by adding clustering information into the model as a variable. For clustering analysis, two types of distances are used between the data-points; the one is Euclidean and the other is 1-corr distance. With these distances, we used the following hierarchical clustering methods: single, complete, average, and Ward (Sibson, 1973; Defays, 1977; Murtagh, 1984; Ward, 1963; Cormack, 1971). In clustering analysis, it is not easy to determine the number of clusters. Different number of clusters could give completely different clustering results. In order to determine the number of clusters more objectively, we use clustering validation measures such as the Silhouette and clustering instability (Rousseeuw, 1987; Fanga and Wangb, 2012). With these measures, the number of clusters is determined. Then, new variables representing the clusters are included in the prediction model in addition to the other clinical variables. For the evaluation of our prediction model, we compare the performance of the model implementing with the pattern information to the model without the pattern information. As expected, we are able to demonstrate that the pattern implemented model outperforms the other model.

2. Subjects and methods

2. 1. Study Subjects

This study participants were 77 individuals with bipolar I (n =31) or II (n =46) disorder who presented depressive symptoms at the time of registration at the Mood Disorders Clinic (MDC) of Seoul National University Bundang Hospital (SNUBH).

Data on demographic and clinical characteristics were obtained from using the case registration form of the MDC and electronic medical records. We collected and analyzed data from the patients who had visited the MDC of SNUBH between January 2005 and July 2011. The data were collected from the systemic prospective follow-up registry of the MDC.

2. 1. 1. Education Status

73 out of 77 BD patients had education information on the first medical examination date (Figure 1).

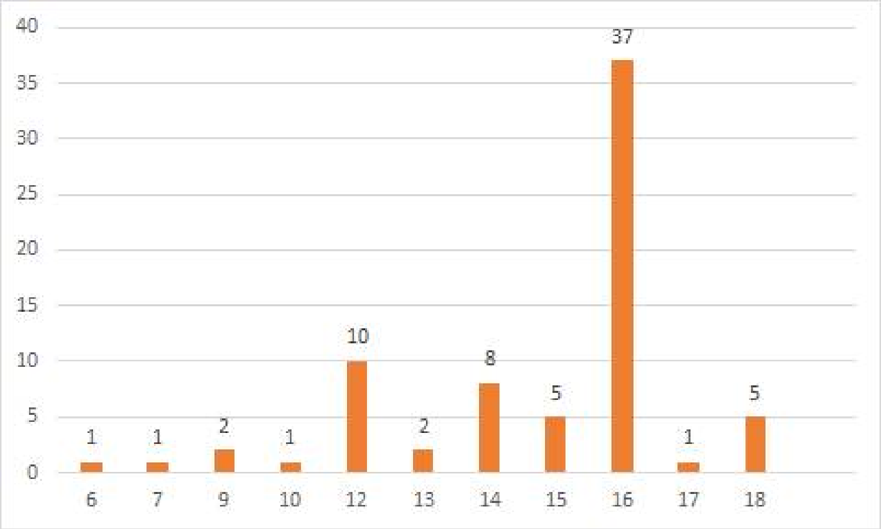

Figure 1. Distribution of Biploar Disorder patients' years of education.

X axis is years of education, Y axis is number of patients. Years of education accounts education after elementary school. To group according to education status, we categorized the data as below (Table 1).

Education Status	Number of Patients
Highschool or below	15
2-year college graduate or current university student	15
Bachelor's degree	37
Master's degree and current student in master's degree	6
Total number of Patients	73

Table 1. Distribution of Biploar Disorder patients' years of education

'Highschool or below' accounts for 12 years or less, '2-year college graduate or current university student' accounts for 13 to 15 years, 'Bachelor's degree' 16 years, and lastly, 17 years of education.

2. 1. 2. Depression counts

There is a report on past depression signs are related to bad prognosis of BD (Treuer and Tohen,2010; Marangell et al, 2009; Tohen et al,1990). Therefore, we have included such information as a main covariate. Out of 77 patients, 62 had past depression information. The plot below is a summary of past depression (Figure 2).

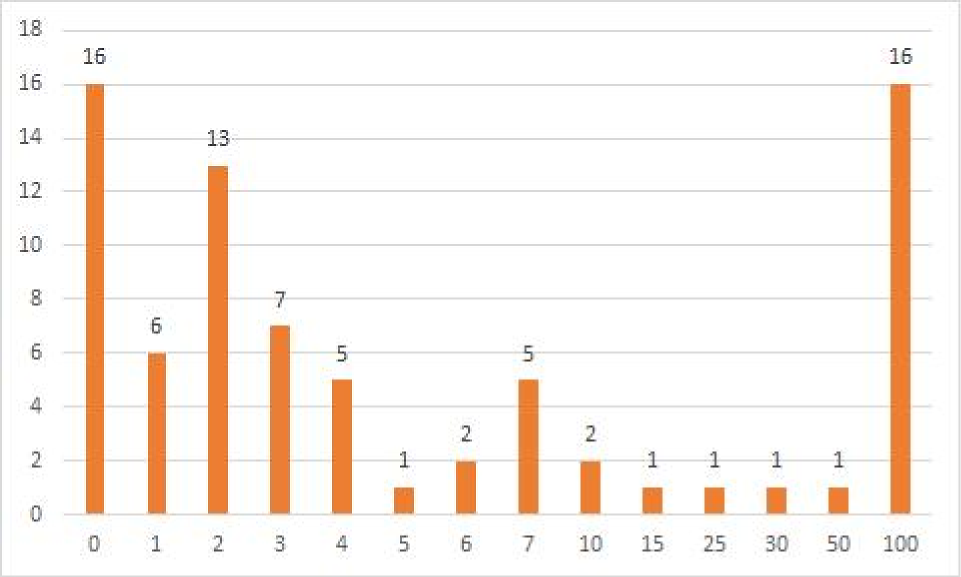

Figure 2. Distribution of bipolar disorder patients' past depression information.

X axis stands for past depression count, Y axis stands for number of patients. Counts more than 100 are defined as 100. The data was categorized using quartile information (Table 2).

Past depression count	number of patients
1 or less	22
2 to 3	20
4 to 15	16
more than 25	19
Patients total	77

Table 2. Distribution of bipolar disorder patients' past depression information.

The cutoff at first quartile is 1, second quartile is 3, third quartile is 15, and fourth quartile is 100.

2. 2. Statistical Method

2. 2. 1 Distance

We use Euclidean distance and 1-corr distance as distances for clustering analysis. First, in n-dimension, Euclidean distance is defined using two points $p=(p_1,p_2,\cdots,p_n)$ and $q=(q_1,q_2,\cdots,q_n)$ on equation (2.3.1).

$$d(p,q)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} \tag{2.3.1}$$

Euclidean distance is known to measure the physical distance between data points, while 1-corr distance is to measure the relatedness between data points by subtracting correlation value of two points from 1 as shown in (2.3.2).

$$1-corr(p,q)=1-\frac{E(pq)-E(p)E(q)}{\sqrt{var(p)}\sqrt{var(q)}} \tag{2.3.2}$$

2. 2. 2. Clustering method

There are several hierarchical clustering methods, and those methods are characterized by their linkage criterion. The linkage criterion determines the distance between sets of observation as a function of the pair-wise distances between all observations: it uses the minimum distance of two observations between-clusters is defined as single (Sibson, 1973). If the maximum distance of between-cluster observations is used, it is defined to be complete (Defays, 1977). If the average of all pair-wise between-cluster observation distances is used, it is the average method (Murtagh, 1984), and Ward's methods is a hierarchical clustering analysis method to minimize the total within-cluster variance, by applying Lance-Williams algorithms (Ward, 1963; Cormack, 1971). In our study, we have used the above four methods.

2. 2. 3. Silhouette

Silhouette is measure to validate the cluster of data. Silhouette is defined as shown in (2.3.3).

$$s(i) = \frac{b(i)-a(i)}{\mathrm{max}a(i),b(i)} = \begin{cases} 1-\frac{a(i)}{b(i)}, \mathrm{if}\ a(i) < b(i) \\ 0, \qquad\quad \mathrm{if}\ a(i) = b(i) \\ \frac{b(i)}{a(i)}-1, \mathrm{if}\ a(i) > b(i) \end{cases} \tag{2.3.3}$$

$a(i)$ is average value of distance in the cluster and $b(i)$ is minimum value of average distances to other clusters. Silhouette measure ranges from -1 to 1 and Silhouette coefficient is defined as the average value of Silhouette values.

2. 2 .4. Selection of the number of clusters via the bootstrap

We define $y = \{y_1, \cdots, y_n\}$ is the dataset with size n, and $\Psi_{y,k}(y)$ is the clustered results. Here, k is the number of clusters. Similarly, bootstrapped sample is defined as y^b, and the clustered results of the bootstrapped samples is $\Psi_{y^b,k}(y)$. The distance between two clustered results is defined as $d(\Psi_{y,k}(y), \Psi_{yb,k}(y))$. The average distance of results from bootstrapping B times, is $s(\Psi,k,n) = \frac{1}{B}\sum_{b=1}^{B} d(\Psi_{y,k}(y), \Psi_{y^b,k}(y))$ and is the definition of instability.

These are the steps of Bootstrap method for selecting the number of clusters:

Step 1. Generate B independent bootstrap sample-pairs $(y, y^b), b = 1, \cdots, B$. Each sample consists of n observations generated from empirical distribution $\hat{\mathrm{F}}$ with replacement.

Step 2. Construct $\Psi_{y,k}(y)$ and $\Psi_{y^b,k}(y)$ based on (y, y^b), $b = 1, \cdots, B$.

Step 3. For each pair, $\Psi_{y,k}(x)$ and $\Psi_{y^b,k}(x)$ calculate their empirical clustering distance

$$d(\Psi_{y,k}(y),\Psi_{y^b,k}(y)) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left| I\{\Psi_{y,k}(y_i)=\Psi_{y,k}(y_j)\} - I\{\Psi_{y^b,k}(y_i)=\Psi_{y^b,k}(y_j)\}\right|$$

Then the clustering instability $s(\Psi,k,n)$ can be estimated by

$$s(\Psi,k,n) = \frac{1}{B}\sum_{b=1}^{B} d(\Psi_{y,k}(y),\Psi_{y^b,k}(y))$$

Step 4. Finally, the optimal number of clusters can be estimated by

$$\hat{k}(n) = \mathrm{arg}min_{2 \le k \le K} s(\Psi,k,n)$$

2. 2. 5. Linear Mixed Effects Model

In our manuscript, we used LMM for prediction of CGI values at given time. Individual differences are considered using random effects in the model. (2.3.4) is a common example of a LMM (Robinson 1991).

$$y = X\beta + Zu + e \tag{2.3.4}$$

where y is a vector of n observable random variables, β is a vector of p unknown parameters having fixed effect, X and Z are known matrices, and u and e are vector of q and n respectively, unobservable random effects such that $E(u)=0, E(e)=0$ and $var\begin{bmatrix}u\\e\end{bmatrix} = \begin{bmatrix}G\,0\\0\;R\end{bmatrix}\sigma^2$ σ^2 where G and R are known positive definite matrices and σ^2 is a positive constant

Fixed effects variables are age, sex, education, time, past depression count, marital status, and bipolar disorder type. Random effects variable was used for patient ID. CGI value is used for dependent variable. In general, LMM(2.3.4)'s parameter is estimated through maximum likelihood and restricted maximum likelihood (Searle, 1992; Vonesh and Chinchilli, 1997). The steps are described in (2.3.5), where θ is assumed to be related to population parameter Σ, and expressed as likelihood function.

$$L(\beta,\theta,\sigma^2|y) = \prod_{i=1}^{n} p(y_i|\beta,\theta,\sigma^2) \tag{2.3.5}$$
$$= \prod_{i=1}^{n} \int p(y_i|\beta,\theta,\sigma^2) p(b_i|\theta,\sigma^2) db_i$$

also,

$$p(y_i|b_i,\beta,\sigma^2) = \frac{\exp(- \parallel y_i - X_i\beta - Z_i b_i \parallel^2 / 2\sigma^2)}{(2\pi\sigma^2)^{n_i/2}} \tag{2.3.6}$$

$$p(b_i|\theta,\sigma^2) = \frac{\exp(-b_i^T \Sigma^{-1} b_i)}{(2\pi)^{q/2}\sqrt{|\Sigma|}} = \frac{\exp(- \parallel \Delta b_i \parallel^2 / 2\sigma^2)}{(2\pi\sigma^2)^{q/2} abs|\Delta|^{-1}} \tag{2.3.7}$$

and Δ satisfies $\frac{\Sigma^{-1}}{1/\sigma^2} = \Delta^T\Delta$. If (2.3.6 , 2.3.7) is applied to (2.3.5), (2.3.8) is derived.

$$L(\beta,\theta,\sigma^2|y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^{n} \parallel \widetilde{y_i} - \widetilde{X_i}\beta - \widetilde{Z_i}\hat{b_i} \parallel^2}{2\sigma^2} \right) \prod_{i=1}^{n} \frac{|\Delta|}{\sqrt{|\widetilde{Z}_{i\}}}} \tag{2.3.8}$$

here, $\widetilde{y_i} = \begin{bmatrix} y_i \\ 0 \end{bmatrix}$, $\widetilde{X_i} = \begin{bmatrix} X_i \\ 0 \end{bmatrix}$, $\widetilde{Z_i} = \begin{bmatrix} Z_i \\ 0 \end{bmatrix}$, $\hat{b_i} = \left(\widetilde{Z_i}^T \widetilde{Z_i}\right)^{-1} \widetilde{Z_i}^T \left(\widetilde{y_i} - \widetilde{X_i}\beta\right)$.

Solving $\hat{\beta}, \hat{\theta}, \hat{\sigma}^2$ that maximizes equation(2.3.4) is the ML method. In LMM, we used AIC measure (Akaike, 1974) to choose the best model.

3. Results

First, we calculated the Euclidean distance and 1-corr distance of the CGI-BP measures, then clustered the results using the following hierarchical clustering methods: single, complete, average, and Ward. To choose the clustering method and its number of clusters (k), we have used Silhouette and clustering instability measure to validate the clusters. The following plots represent the results from such process (Figure 3, Figure 4).

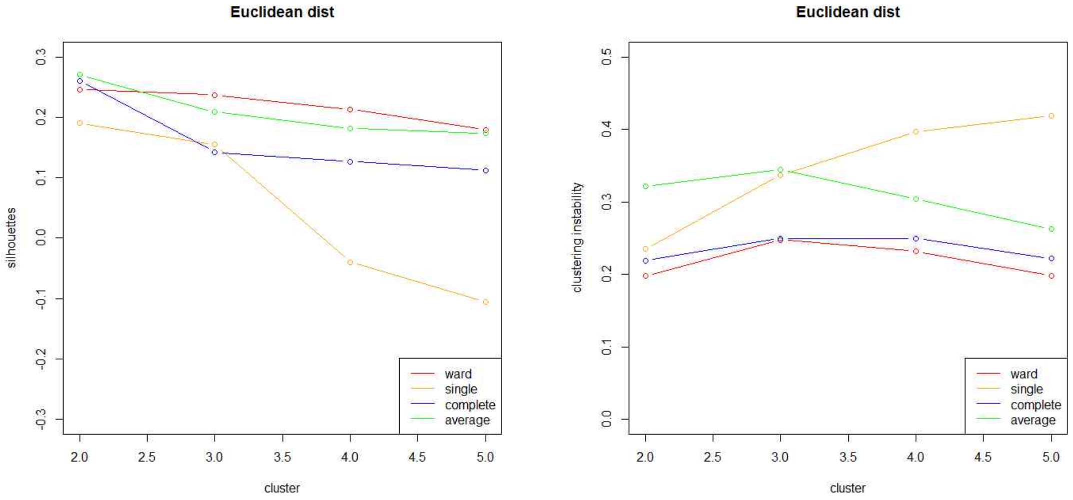

Figure 3. Silhouette and clustering instability output with Euclidean distance

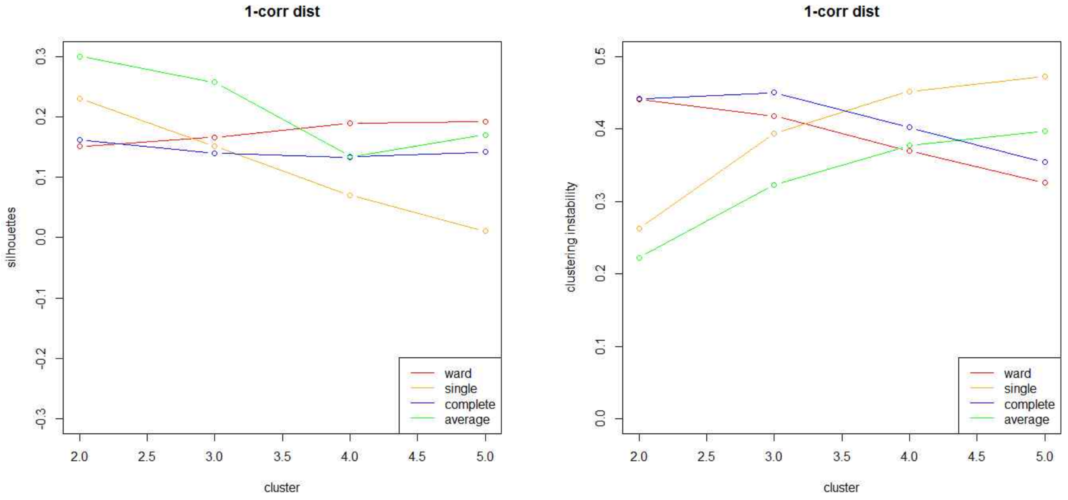

Figure 4. Silhouette and clustering instability output with 1-corr distance

A guide to interpreting the Silhouette and clustering instability measures suggest a greater Silhouette value or a smaller clustering instability value denotes a better clustered output. When using Euclidean distance, Figure 3 shows both Silhouette and clustering instability suggests $k=2$ as best, while Silhouette suggested average, complete, and Ward were compatible, while clustering instability suggested only complete and Ward kept their performances in several k's.

In Figure 4, when using the 1-corr distance, also Silhouette and clustering instability showed $k=2$ to outperform other k's. Average method proved to be the best in both criteria. To check how the CGI-BP measures have been clustered, we used the scatter plot below.

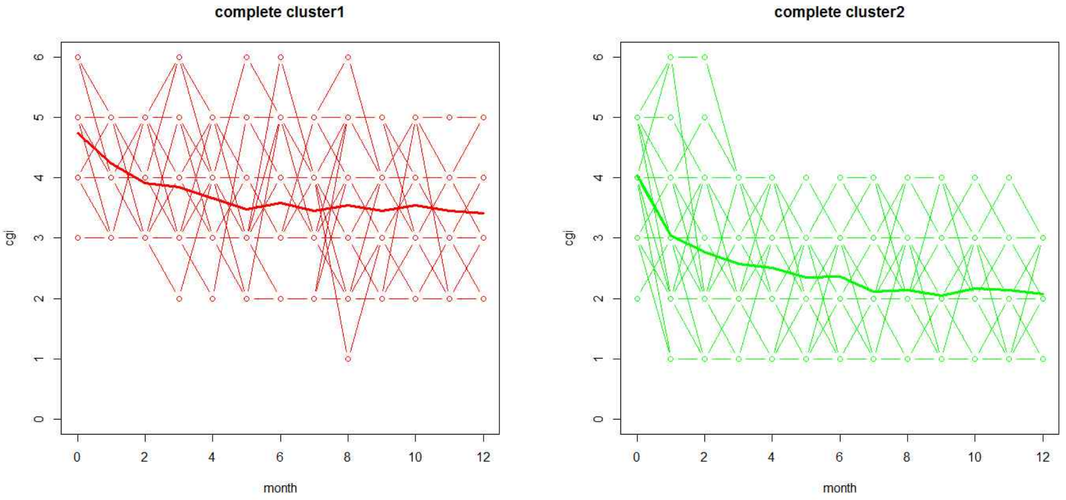

Figure 5. Scatter plot of each plot using complete method

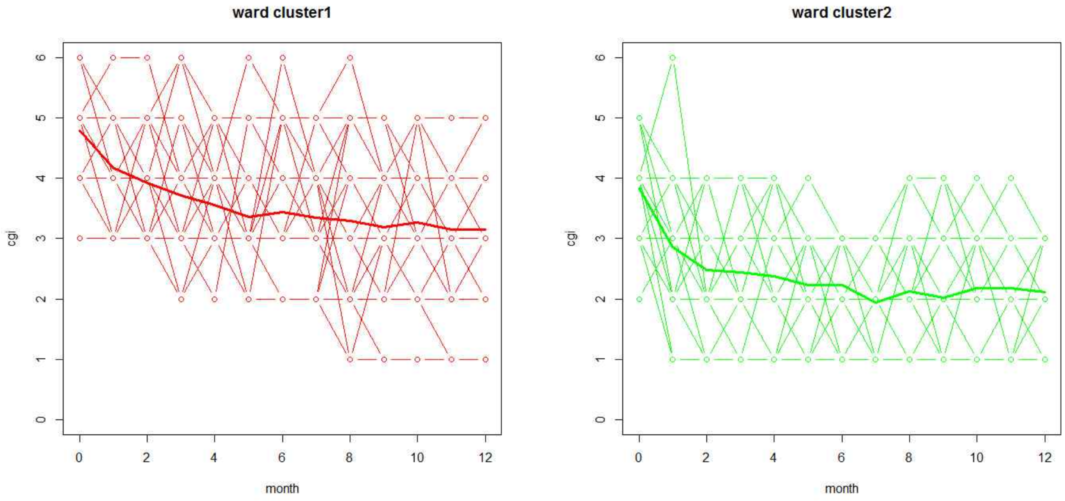

Figure 6. Scatter plot of each plot using Ward method

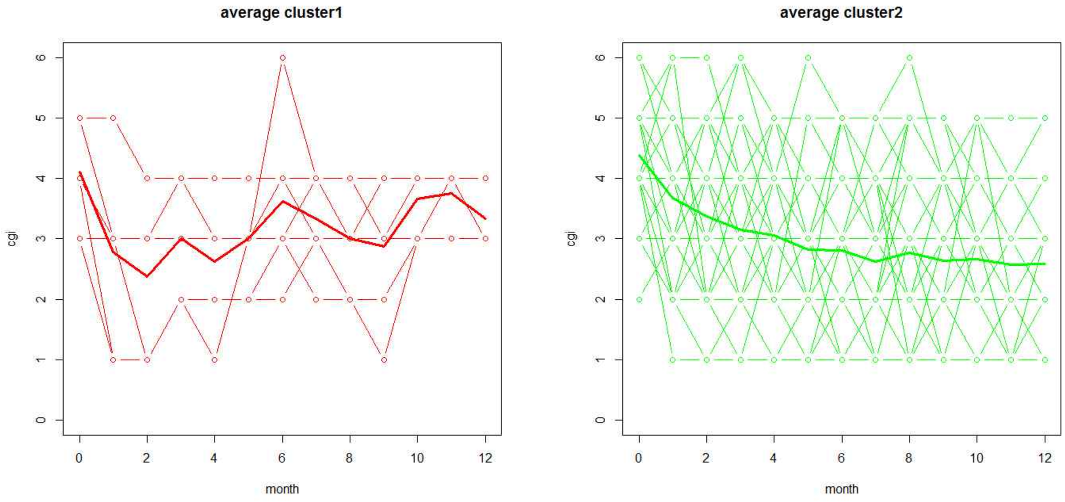

Figure 7. Scatter plot of each plot using average method

The ‘bold line’ of each scatter plot is a graph of broken line of average value at each time points. Figure 5 suggests the first cluster’s CGI-BP average value stays above 3; the patient is not showing improvement, while the second cluster values seem to drop from a high value; this group of patients are showing positive progress. Similarly in Figure 6, the first cluster seems to show no improvements, while the second cluster shows a drop in CGI-BP average; the patients are improving. Apart from the previous two figures, Figure 7, shows an unstable wobble of CGI-BP measures for the first cluster, and the second cluster shows a slight drop and values being leveled off. Such plot may suggest the status of first clustered group shows instability, while the individuals in second group found stability in their lives.

We have compared the models with and without clustered group information to gain information if using the cluster group information increases the prediction power, via AIC and prediction error. First, we used the all covariates: age, sex, education, time, and marital status in the LMM models. Then, using the AIC measures, we chose the models with lowest AIC (Table 3).

model	AIC
$y_{ij} = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 disease + \beta_4 married + \beta_5 dep$ $+ \beta_6 edu + \beta_7 time + b_{0i} + \epsilon_{ij}$	665.54
$y_{ij} = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 disease + \beta_4 married + \beta_5 dep + \beta_6 time + b_{0i} + \epsilon_{ij}$	661.38
$y_{ij} = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 married + \beta_4 dep + \beta_5 time + b_{0i} + \epsilon_{ij}$	659.52
$y_{ij} = \beta_0 + \beta_1 age + \beta_2 married + \beta_3 dep + \beta_4 time + b_{0i} + \epsilon_{ij}$	658.47
$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + b_{0i} + \epsilon_{ij}$	656.49
$y_{ij} = \beta_0 + \beta_1 dep + \beta_2 time + b_{0i} + \epsilon_{ij}$	658.57

Table 3. Linear Mixed Effects Models without group information.

From the pool of models, we chose the following model,

$$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + b_{0i} + \epsilon_{ij} \tag{3.1}$$

Here, b_{0i} is random effect, we added group and also interaction term between group and time in equation (3.2) and (3.3), respectively.

$$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp + b_{0i} + \epsilon_{ij} \tag{3.2}$$

$$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp + \beta_5 grp * time + b_{0i} + \epsilon_{ij} \tag{3.3}$$

If we use the group informations in the models, the output is as follows (Table 4).

clustering method	model	AIC
complete	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp + b_{0i} + \epsilon_{ij}$	590.94
	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp$ $+ \beta_5 grp^* time + b_{0i} + \epsilon_{ij}$	571.60
ward	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp + b_{0i} + \epsilon_{ij}$	606.09
	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp$ $+ \beta_5 grp^* time + b_{0i} + \epsilon_{ij}$	607.74
average	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp + b_{0i} + \epsilon_{ij}$	656.80
	$y_{ij} = \beta_0 + \beta_1 married + \beta_2 dep + \beta_3 time + \beta_4 grp$ $+ \beta_5 grp^* time + b_{0i} + \epsilon_{ij}$	603.98

Table 4. Linear Mixed Effects Models with group information.

Comparing the AIC values of models with clustered group information versus the models without group information, we observed all methods without average methods showed lower AIC values for the models with clustered group information. Using the complete methods with interaction in the model showed the lowest AIC value.

Prediction error value is derived from the prediction of 13th time point value from first 12 time point data, it is the squared value of the difference between real 13th data and the predicted value. The results from the validation of above clustering outputs and validation with 12 time point data only are plotted as followed (Figure 8, Figure 9).

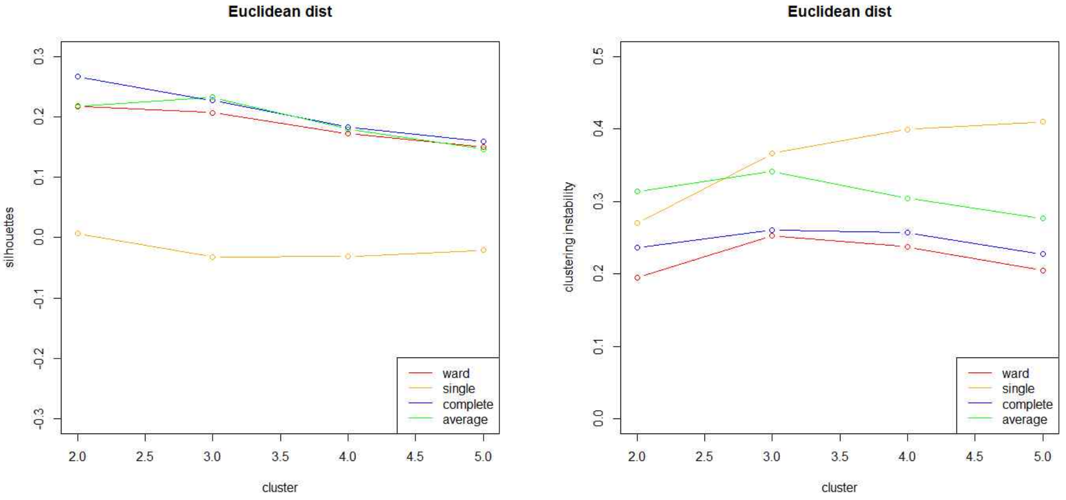

Figure 8. Silhouette and clustering instability output with Euclidean distance using 12 time point

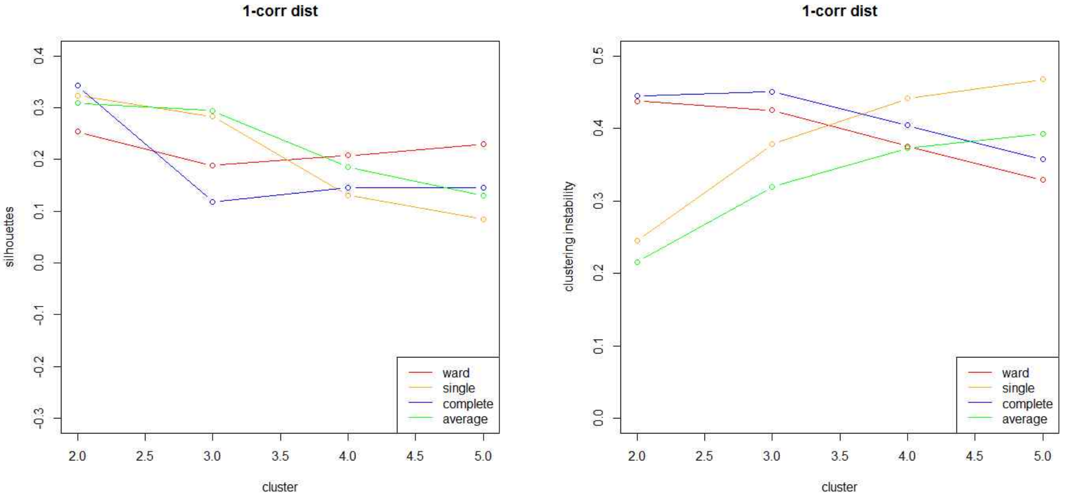

Figure 9. Silhouette and clustering instability output with 1-corr distance using 12 time point

From Figure 8 and Figure 9, we could infer using 2 clusters works best for both Euclidean and 1-corr distance, but complete and ward works well for the former and average method worked well for the latter.
The following models are similar to the aforementioned Table 1 and Table 2, but using data from 12 time points (Table 5).

model	AIC
$y_{ij}=\beta_0+\beta_1 age+\beta_2 sex+\beta_3 disease+\beta_4 married+\beta_5 dep$ $+\beta_6 edu+\beta_7 time+b_{0i}+\epsilon_{ij}$	602.49
$y_{ij}=\beta_0+\beta_1 age+\beta_2 sex+\beta_3 disease+\beta_4 married+\beta_5 dep+\beta_6 time+b_{0i}+\epsilon_{ij}$	598.25
$y_{ij}=\beta_0+\beta_1 age+\beta_2 sex+\beta_3 married+\beta_4 dep+\beta_5 time+b_{0i}+\epsilon_{ij}$	596.37
$y_{ij}=\beta_0+\beta_1 age+\beta_2 married+\beta_3 dep+\beta_4 time+b_{0i}+\epsilon_{ij}$	594.81
$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+b_{0i}+\epsilon_{ij}$	593.53
$y_{ij}=\beta_0+\beta_1 dep+\beta_2 time+b_{0i}+\epsilon_{ij}$	595.89

Table 5. Linear Mixed Effects Models without group information. using 12 time point

As a result, we chose the above model with prediction error 60.07. We then added group and interaction term between time and group to get the following prediction error values.

clustering method	model	prediction error
complete	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp+b_{0i}+\epsilon_{ij}$	59.79
	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp$ $+\beta_5 grp*time+b_{0i}+\epsilon_{ij}$	57.43
ward	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp+b_{0i}+\epsilon_{ij}$	60.68
	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp$ $+\beta_5 grp*time+b_{0i}+\epsilon_{ij}$	60.58
average	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp+b_{0i}+\epsilon_{ij}$	60.03
	$y_{ij}=\beta_0+\beta_1 married+\beta_2 dep+\beta_3 time+\beta_4 grp$ $+\beta_5 grp*time+b_{0i}+\epsilon_{ij}$	60.95

Table 6. Linear Mixed Effects Models with group information. prediction error using 12 time point

Comparing the prediction error values of the models with and without group information, we observed the models with group information having lower prediction error except average method. In conclusion, complete method with interaction term is best for model with group information; it is chosen from prediction error and AIC criterion.

4. Discussion

To summarize, we checked if the models with clustered group information predicted CGI-BP measures well. We have already concluded DEP and marital status plays a key role in CGI prediction, and used the Euclidean and 1-corr distance on four hierarchical clustering. The clustered sets were measured using Silhouette and clustering instability, which proved using 2 clusters worked best. For Euclidean distance, complete and Ward method worked best, and for 1-corr distance, average method was chosen at best. Using the clustered information from the above methods, we have fitted the Linear Mixed Effects Model and computed corresponding AIC measures. Models except for average method showed lower AIC values when group information is not used, but when the group information is used, we could observe the complete method with interaction showed the lowest AIC value. We could observe the same result using prediction error criterion. Therefore, we chose the model with 2 clusters, complete method, and using group and interaction information. We can use this model to predict patient status, and apply suitable treatment to patients.

The missing rate in CGI-BP data is high, and thus experienced some problem in making the prediction mode, and further study should focus on solving the problem statistically to make a better prediction model. Our study is noteworthy as we made a prediction model for future patient status by using CGI-BP measures. If a more precise model can be made in further studies, the predicted values can be used to treat high-risk patients separately.

5. Reference

Judd, L.L., Akiskal, H.S., Schettler, P.J., Endicott, J., Maser, J., Solomon, D.A., Leon, A.C., Rice, J.A., Keller, M.B. (2002), The long-term natural history of the weekly symptomatic status of bipolar I disorder. Archives of General Psychiatry 59, 530 – 537.

Marangell, L.B., Dennehy, E.B., Miyahara, S., Wisniewski, S.R., Bauer, M.S., Rapaport, M.H., Allen, M.H. (2009), The functional impact of subsyndromal depressive symptoms in bipolar disorder: data from STEP-BD. Journal of Affective Disorders 114, 58 – 67.

Tohen, M., Hennen, J., Zarate Jr., C.M., Baldessarini, R.J., Strakowski, S.M., Stoll, A.L., Faedda, G.L., Suppes, T., Gebre-Medhin, P., Cohen, B.M. (2000), Two-year syndromal and functional recovery in 219 cases of firstepisode major affective disorder with psychotic features. The American Journal of Psychiatry 157, 220 – 228.

Peter J. Rousseeuw (1987), Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 53 – 65. doi:10.1016/0377-0427(87)90125-7.

Bauer, M., Juckel, G., Correll, C.U., Leopold, K., Pfennig, A. (2008), Diagnosis and treatment in the early illness phase of bipolar disorders. European Archives of Psychiatry and Clinical Neuroscience 258 (Suppl 5), 50 – 54.

Swann, A.C. (2005), Long-term treatment in bipolar disorder. The Journal of Clinical Psychiatry 66 (Suppl 1), 7 – 12.

Moon E, Chang JS, Kim MY, Seo MH, Cha B, Ha TH, Choi S, Cho HS, Park T, Ha K. (2012). Dropout rate and associated factors in patients with bipolar disorders. Journal of Affective Disorders. 2012 Dec 1;141(1):47-54.

Cormack, R. M. (1971), A Review of Classification, Journal of the Royal Statistical Society, Series A, 134(3), 321–367.

Ward, J. H., Jr. (1963), Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58, 236–244.

Yixin Fanga and Junhui Wangb. (2012), Selection of the number of clusters via the bootstrap method Computational Statistics & Data Analysis Volume 56, Issue 3, 1 March 2012, Pages 468–477

Searle, S.R., Casella, G. and McCulloch, C. E. (1992), Variance components , Wiley Online Library.

Vonesh, E. F. and Chinchilli, V. M. (1997), Linear and nonlinear models for the analysis of repeated measurements, Vol.154, CRC Press.

Robinson, G.K. (1991), That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science 6 (1): 15–32. doi:10.1214/ss/1177011926. JSTOR 2245695.

Akaike, H. (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19(6): p. 716–723.

Joan Busner and Steven D. Targum (2007), MD The Clinical Global Impressions Scale Psychiatry (Edgmont)., July; 4(7): 28–37.

Avital Cnaan, Nan M. Laird and Peter Slasor. (1977), TUTORIAL IN BIOSTATISTICS USING THE GENERAL LINEAR MIXED MODEL TO ANALYSE UNBALANCED REPEATED MEASURES AND LONGITUDINAL DATA. STATISTICS IN MEDICINE, VOL. 16, 2349Ð2380

R. Sibson (1973), SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (British Computer Society) 16 (1): 30–34.

D. Defays (1977). An efficient algorithm for a complete link method. The Computer Journal (British Computer Society) 20 (4): 364 - 366.

Murtagh F (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. Computational Statistics Quarterly 1: 101 - 113.

국문초록

많은 의학 연구에서 흔히 반복측정 자료가 생산된다. 반복 측정된 자료 분석의 대표적인 목적 중 하나는 관측된 자료를 이용하여 미지의 값을 예측하는 것이다. 반복 측정된 자료는 시간에 대해 산점도을 그렸을 때 특정한 패턴이 보이는 경우가 종종 있다. 이 논문에서는 반복 측정된 자료 분석에서 패턴정보를 값을 이용하면 예측력을 높일 수 있다는 것을 증명했다. 우리는 자료의 패턴을 clustering 방법으로 clustering하고 clustering 정보를 모형에 변수로 추가하여 예측모형을 만드는 것을 제안한다. 우리가 제안한 방법을 bipolar 환자들의 실제 자료를 이용해 확인했다. 환자의 우울증 진행경과를 예측하기 위하여 Clinical Global Impression (CGI)값을 이용한다. 먼저 CGI값 사이의 유클리디안 거리와 1-corr 거리를 계산하여 계층적 clustering 방법으로 clustering을 하여 가장 좋은 결과를 선택한다. 여기에서 구해지는 clustering 정보를 사용하여 가장 좋은 우울증 진행경과 예측모형을 만든다. 자료의 개인의 효과를 고려하기 위하여 선형 혼합 효과 모형을 사용하고 랜덤효과를 고려한 예측모형을 만든다. AIC기준과 예측오차로 비교했을 때 우리가 제안한 모델이 clustering 정보를 쓰지 않은 모델보다 좋은 모델이고 예측력이 더 좋다고 할 수 있다.

……………………………………

주요어 : clustering 방법, 선형 혼합 효과 모형, 예측모형

학　번 : 2012-20224