The Optimized K-Means Clustering Algorithms To Analyzed the Budget Revenue Expenditure in Padang

Dony Novaliendry*  
Departement of Electronic Engineering, National Kaohsiung University of Applied Sciences  
Kaohsiung, Taiwan  
dony_novaliendry@ymail.com

Yeka Hendriyani  
Departement of Electrical Engineering, Technical Faculty, State University of Padang  
Padang, Indonesia  
yecha_053@yahoo.co.id

Cheng-Hong Yang**  
Departement of Electronic Engineering, National Kaohsiung University of Applied Sciences  
Kaohsiung, Taiwan  
chyang@cc.kuas.edu.tw

Hafilah Hamimi  
Departement of Electrical Engineering, Technical Faculty, State University of Padang  
Padang, Indonesia  
hafilahhamimi03@gmail.com

Abstract—APBD is a systematic detailed list of receipts, expenditures and local spending within a year arranged in PERMENDAGRI No. 16 of 2006, so that the data of APBD can be used as guidelines for governments and local expenditures in carrying out activities to raise revenue to maintain economic stability and to avoid inflation and deflation. Government financial institutions in areas such as DPKA Padang, experienced difficulties in identifying the relevance of each archive data on APBD, that result in a data warehouse. In addition, to the administration, APBD in the government of Padang have not been effective. To minimize the difficulty in identifying data archive of APBD, then the data warehouse can be used to produce knowledge using the techniques of Data Mining (DM). The method that is used are clustering and forecasting. Clustering performed using the K-Means Algorithm while forecasting is done by using multiple linear regressions. These methods intended to classify and identify the data in the budget that have certain characteristics in common, and can predict the value of APBD for the following years.

Keywords—Clustering, Multiple Linear Regression, K-Means, Forecasting

Introduction

The fact that there are huge numbers of data in the institution which manages local budget has caused the difficulty in organizing these data. The management of Local Government Budget (APBD) in the city government of Padang has not been effectively implemented as regulated in PERMENDAGRI (Regulations from Internal Affairs Minister) No. 17 Year 2007. This ineffective implementation of the local government budget management in Padang is caused by many problems in its management itself; one of them is inadequate human resources.

"The data of APBD is an annual budget planning of local government which is approved by Regional Representative Assembly (DPRD)” (Law No.17 Year 2003 article 1 paragraph 8 about State Budget) which is administered by the Office of Financial Management and Asset (DPKA) in that city.

APBD is a systematic list of revenue, expenditure, and local expense within a certain period of time (1 year) which is regulated in PERMENDAGRI No. 16 Year 2006. Thus, APBD can be used as guidelines for government and local expenses in conducting activities for increasing the revenue in order to maintain the economic stability, and avoid inflation and deflation.

The data of APBD have been grouped into revenue, expense, and financing. The following are the data of local government budget in Padang which consist of Local Own-source Revenue, Transfer Revenue, and other legitimate Revenue.

<table>
<thead>
<tr>
<th>APBD</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Own-source Revenue</td>
<td>133.164</td>
<td>120.926</td>
<td>164.935</td>
<td>187.627</td>
<td>238.889</td>
</tr>
<tr>
<td>Transfer Revenue</td>
<td>768.005</td>
<td>841.254</td>
<td>1.059.022</td>
<td>1.258.375</td>
<td>1.488.658</td>
</tr>
<tr>
<td>Legitimate Revenue</td>
<td>2.335</td>
<td>71.782</td>
<td>56.274</td>
<td>6.054</td>
<td>18.183</td>
</tr>
<tr>
<td>Expense</td>
<td>1.050.192</td>
<td>1.185.934</td>
<td>1.321.830</td>
<td>1.533.282</td>
<td>1.875.768</td>
</tr>
</tbody>
</table>
The problem faced by local financial institution in Padang is how to organize and identify this huge number of APBD data. However, along with the development of Information and Technology (IT), there are several ways and solutions emerged in order to solve or minimize this problem. One of them is the use of clustering and forecasting methods of Data Mining (DM). These methods are intended to group and identify APBD data which are similar in certain characteristics, and predict the value of APBD in the future.

Data mining technique can process the big capacity of data information extraction. “Data Mining is a process of seeking pattern, and hidden relations in big capacity data in order to classify, estimate, forecast, associate rules, sequential pattern, clustering, regression, describe, and visualize”[1]. The method used to group APBD data is clustering for it can group and identify APBD data which are similar in certain characteristics; meanwhile, forecasting with multiple linear regressions is used to predict the future APBD, and see the relations of each budget.

According to Prasetyo [2], “among several methods of clustering, algorithm k-means is the most simple and common to be implemented”. Furthermore, Agusta (2005) [2] states, “K-Means is one of non-hierarchy clustering methods which is able to divide data into cluster, so that data with similar characteristics are grouped into one similar cluster, and the different ones are grouped into another cluster”.

By implementing algorithm k-means in the process of APBD clustering, we can group and determine the number of adequate cluster, determine the parameter of borders based on the characteristics in each cluster so that we can predict the total of local budget in the future by using multiple linear regression in APBD Padang.

**Methods**

**Data Mining**

Data mining is a term used to find the hidden information in a database. David Hand, Heikki Mannila, and Phadaraic Smyth from MIT in Prabowo, Rahmadya et al.[3] state that “Data mining is an analysis of data (commonly high-capacity data) in order to see clear relations and conclude things which are previously unknown by using a current and beneficial way for the owner of the data themselves”.

Larose (2006) [2] states that data mining can be classified into six categories, namely: (a) description, (b) estimation, (c) prediction, (d) classification, (e) clusterization, and (f) association.

**Clustering**

Clustering is a discovery tool that is used to reveal the relationships and structures of data from something that was not previously clear into useful knowledge when discovered.

The main goal of clustering method is grouping a number of data/objects into clusters so that in each cluster will contain similar data.

**K-Means Algorithm**

K-Means algorithm is the most popular and widely used algorithm in clustering that is used in the industrial world. K points from the data set as the initial cluster center, putting the sample to the class where the nearest cluster center in [4][5]. Calculating data object in each new formed cluster to get the mean center of a new cluster, if two adjacent cluster centers without any change, then the sample adjustment is completed, and clustering criterion function has converged. The character of the algorithm is to investigate the classification of the sample correct or not in every iteration. If all samples were classified correctly, it will not in every iteration. If all samples were classified correctly, it will not adjust, which marks convergence, then the algorithm finish; otherwise it needs to be adjusted into the next iteration.

**Forecasting**

Forecasting is a process for making a statement on an incident in which the incident is not known or observable[6]. A common practice in forecasting is to estimate the expected value of a variable to be studied in the future.

**Multiple Linear Regression**

Regression equation with many variables X, ie variable influencers. Where have k variables influencers, the regression line equation becomes:

\[\hat{Y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k\]  

(1)

To find a to b, the following equation :

\[na + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \cdots + b_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i\]

\[a \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + b_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i\]  

(2)

**WEKA Data Mining Tool**

WEKA is a data mining applications based on open source (GPL) and Java engine [7].

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WEKA is free software available under the GNU General Public License. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [8].

WEKA is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be
applied directly to a dataset or called from your own Java code [9].

WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection [10]. All of WEKA’s techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using WEKA. Another important area that is currently not covered by the algorithms included in the WEKA distribution is sequence modeling [8].

Data Analysis Techniques

Technique of K-Means Clustering

“K-Means clustering” is a partitioning method. The function k-means partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. K-means clustering operates on actual observations and creates clusters. The k-means clustering is often more suitable for large data [11]. The simple K-means is a common unsupervised classification method. Its effectiveness depends on accurate estimation of k clusters for each spectral data [12].

Data clustering techniques using the K-Means algorithm:

a) Determine the number of clusters k.
b) The initialization of k cluster centers can be done randomly and used as initial cluster centers.
c) Allocate all data/objects to the nearest cluster. The proximity of the two objects is determined by the distance of the object. To calculate the distance of all the data to the cluster center point using the Euclidean distance theory formulated as follows:

\[ D_{(x,y)} = \sqrt{(X_{1x} - x_{1y})^2 + (X_{2x} - x_{2y})^2 + \ldots + (X_{nx} - x_{ny})^2} \] (3)

whereas:

- \( D_{(x,y)} \) = distance data of \( x \) to the cluster center of \( y \)
- \( X_{nx} \) = i-data on attribute data of \( k \)
- \( X_{ny} \) = the center point of \( j \) on the attribute of \( k \)
d) The shortest distance between the center of the cluster with the data/object determines the cluster position of a data/object
e) Calculate the ratio between “Between Cluster Variation (BCV)” and “Within Cluster Variation (WCV).”

\[ BCV = d(m_1,m_2) + d(m_1,m_3) + d(m_2,m_3) \] (4)

\[ WCV = \frac{BCV}{n} \] (5)

Then compare BCV with WCV. If the ratio is enlarged, continue to process f, otherwise stop the process and continue to process h.
f) Recalculate the center of the new cluster membership. This is computed by determining the centroid/center cluster.

\[ C(i) = \frac{x_1 + x_2 + x_3 + \ldots}{\sum x} \] (6)
g) Assign each object to put on the new cluster center, if the center of the cluster changed, back to c, otherwise clustering is complete.
h) Analyze the results.
i) Stages of analyzing the results obtained in the clustering process.
j) The process is done.

Forecasting Technique Using Multiple Regression Linear

Four Padang City budget data attributes that is used: the data revenue (X1), the data transfer income (X2), the data other legitimate income (X3), and the data amount of the total budget (Y). Calculate the gradient b1, b2, b3 and constant by using the following equation:

\[ n a + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + b_3 \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_i \]

\[ a \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{2i} x_{1i} + b_3 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{2i} y_i \]

\[ a \sum_{i=1}^{n} x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 + b_3 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} x_{2i} y_i \]

\[ a \sum_{i=1}^{n} x_{3i} + b_3 \sum_{i=1}^{n} x_{3i} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} x_{3i} + b_3 \sum_{i=1}^{n} x_{3i}^2 = \sum_{i=1}^{n} x_{2i} y_i \]

The results of the K-Means clustering

Table II. APBD of Padang (2009-2013) before clustering

<table>
<thead>
<tr>
<th>Year</th>
<th>Local Revenue</th>
<th>Transfer Revenue</th>
<th>Legitimate Revenue</th>
<th>Expense</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>133.164</td>
<td>768.005</td>
<td>2.335</td>
<td>1.050.192</td>
</tr>
<tr>
<td>2010</td>
<td>120.926</td>
<td>841.254</td>
<td>71.782</td>
<td>1.185.934</td>
</tr>
<tr>
<td>2011</td>
<td>164.935</td>
<td>1.059.022</td>
<td>56.274</td>
<td>1.321.830</td>
</tr>
<tr>
<td>2012</td>
<td>187.627</td>
<td>1.258.375</td>
<td>6.054</td>
<td>1.533.282</td>
</tr>
<tr>
<td>2013</td>
<td>238.889</td>
<td>1.488.658</td>
<td>18.183</td>
<td>1.875.768</td>
</tr>
</tbody>
</table>

Data member value of APBD (2009-2013) after clustering are as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>Local Revenue</th>
<th>Transfer Revenue</th>
<th>Legitimate Revenue</th>
<th>Expense</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>133.164</td>
<td>768.005</td>
<td>2.335</td>
<td>1.050.192</td>
</tr>
<tr>
<td>2010</td>
<td>120.926</td>
<td>841.254</td>
<td>71.782</td>
<td>1.185.934</td>
</tr>
<tr>
<td>2011</td>
<td>164.935</td>
<td>1.059.022</td>
<td>56.274</td>
<td>1.321.830</td>
</tr>
<tr>
<td>2012</td>
<td>187.627</td>
<td>1.258.375</td>
<td>6.054</td>
<td>1.533.282</td>
</tr>
<tr>
<td>2013</td>
<td>238.889</td>
<td>1.488.658</td>
<td>18.183</td>
<td>1.875.768</td>
</tr>
</tbody>
</table>

Table III. Member data value of APBD on 1st Cluster

<table>
<thead>
<tr>
<th>Year</th>
<th>Local Revenue</th>
<th>Transfer Revenue</th>
<th>Legitimate Revenue</th>
<th>Expense</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>133.164</td>
<td>768.005</td>
<td>2.335</td>
<td>1.050.192</td>
</tr>
<tr>
<td>2010</td>
<td>120.926</td>
<td>841.254</td>
<td>71.782</td>
<td>1.185.934</td>
</tr>
<tr>
<td>2011</td>
<td>164.935</td>
<td>1.059.022</td>
<td>56.274</td>
<td>1.321.830</td>
</tr>
<tr>
<td>2012</td>
<td>187.627</td>
<td>1.258.375</td>
<td>6.054</td>
<td>1.533.282</td>
</tr>
<tr>
<td>2013</td>
<td>238.889</td>
<td>1.488.658</td>
<td>18.183</td>
<td>1.875.768</td>
</tr>
</tbody>
</table>

Result
Table IV Member data value of APBD on 2nd Cluster

<table>
<thead>
<tr>
<th>Year</th>
<th>Local Revenue</th>
<th>Transfer Revenue</th>
<th>Legitimate Revenue</th>
<th>Expense</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>120.926</td>
<td>841.254</td>
<td>71.782</td>
<td>1.185.934</td>
</tr>
<tr>
<td>2011</td>
<td>164.935</td>
<td>1.059.022</td>
<td>56.274</td>
<td>1.321.830</td>
</tr>
</tbody>
</table>

Table V Member data value of APBD on 3rd Cluster

<table>
<thead>
<tr>
<th>Year</th>
<th>Local Revenue</th>
<th>Transfer Revenue</th>
<th>Legitimate Revenue</th>
<th>Expense</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>187.627</td>
<td>1.258.375</td>
<td>6.054</td>
<td>1.533.282</td>
</tr>
<tr>
<td>2013</td>
<td>238.889</td>
<td>1.488.658</td>
<td>18.183</td>
<td>1.875.768</td>
</tr>
</tbody>
</table>

After the entire value of the cluster members is formed it will be known that which cluster is the local revenue, transfers revenue, legitimate revenue and the total expenditure.

The results of the multiple linear regression forecasting

The process of multiple linear regression for forecasting of the future budget has resulted prediction by the following equation:

\[ Y = 171.551.3 + 1.791 (X_1) + 0.825 (X_2) + 0.836 (X_3) \]

whereas \( X_1 = 293.573 \), \( X_2 = 1.667.561 \) dan \( X_3 = 334.723 \) (in million), so it can be predicted total state budget for 2014 is:

\[
Y = 171.551.3 + 1.791 (X_1) + 0.825 (X_2) + 0.836 (X_3) \\
Y = 171.551.3 + 1.791 (293.573) + 0.825 (1.667.561) + 0.836 (334.723) \\
Y = 171.551.3 + 2525.789.2 + 1.375.737.825 + 279.828.4 \\
Y = 2.352.906.8
\]

The total budget in 2014 was estimated at \( 2,352,906.8 \) million.

Testing Clustering with WEKA

Figure 1 is the result of Visualize Cluster Assignments of WEKA tool. It shows a graph of the results of clustering with k-means algorithm. Whereas Figure 2 is a viewer ARFF of WEKA tool, the results of clustering with WEKA tool.
From several researches collected, it can be concluded that:

Firdausi et al. in 2011 in his research An Analysis of Financial Distress by Using Data Mining Approach in Go-Public Manufacture Industry in Indonesia [12] explains about the comparison between algorithm K-Means and Fuzzy C-Means (FCM). It was found that Algorithm K-Means worked better than algorithm FCM in grouping financial data of company which will be bankrupt. The formulas used in this research were sum squared error (SSE) and icdrate (internal cluster dispersion rate). The finding of this research showed that the smallest value of SSE was in K-Means which indicated that the total error of grouping by using this method was small. Thus, it can be concluded that K-means method had the smallest value of error and better than FCM method.

By using FCM method, the diversity in clusters (Sum of Squared Within) SSW had the highest value; meanwhile, the diversity between clusters (Sum of Squared Between) SSB was the lowest. Therefore, FCM had the highest value of icdrate compared to K-means. The high value of SSW indicated that there were many different data in a cluster which were grouped by FCM method. Furthermore, the low value of SSB indicated that the data in cluster 1 and cluster 2 were not so much different. Meanwhile, the lowest score in K-Means indicated that the diversity in SSW data was low and the diversity in SSB data was high. Hence, concerning the value of SSE and icdrate by using K-Means and FCM, K-Means was found better.

Dwi Novianti Nangi in 2012 in her journal The Implementation of Algorithm K-Means for clustering Local Government Budget in XYZ Regency conveys that algorithm K-Means was good to be used to find out the number of cluster [14]. In this research, she explains how to determine the most accurate number of cluster, and how to develop and analyze cluster prototype in order to determine parameter of borders regarding the characteristics of each cluster. It was concluded that algorithm K-Means was good to group a large number of data such as financial data of a government institution.

Wahyuni (2009) in her journal entitled The Use of Cluster-Based Sampling to Understand Multi Objective Association Rules states that algorithm K-Means is better than FCM [15]. Her writing explained the comparison of these two methods, K-Means and FCM. The finding revealed that the understanding of multi objective association rules by using samples which had been previously clustered had a better result of association rules. It was indicated from its average
score which was higher than those which were not clustered. Furthermore, K-Means method was better than FCM. It can be seen from the average score of confidence which used K-Means method was higher than that which used FCM method.

Conclusions

This research analyzed the data of local government budget by using clustering with algorithm K-Means, and forecasting with multiple linier regressions. This research used 4 attributes; they are Local Own-Source Revenue (PAD), Transfer Revenue (PT), other Legitimate Revenue (PYS), and the total of Expense. In clustering, the examining of 1 parameter of centroid value is 3 centroid values.

The clustering of APBD (2009-2013) by using clustering k-means was resulted on three clusters. The result of forecasting of APBD in Padang based on the APBD of last five years (2009 -2013) and the relations among attributes by using multiple linier regression was 2.353.906,8 in 2014.

ACKNOWLEDGMENT

Thanks to DIKTI and the Office of Local Financial Management of Padang, Indonesia.

REFERENCES