Implementation of EM Algorithm in Data Mining for Clustering Female Cooperative

Putu Angelina Widya[1] and Made Sudarma[2]

[1,2]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
*Email : angelina.widya@student.unud.ac.id, msudarma@unud.ac.id.

***Abstract*— Cooperatives have a key role for women by providing access to generating income activities. Women have access to general production resources like credit, land, marketing facilities, infrastructure, equipment, and technology. All of that resources can increase their income. By shaping themselves into cooperatives, they can get benefit from economies scale and increase their access to the labor market. Cooperatives can be the best instruments to improve women's welfare and for the development of transformations such as more open minded by using the most profitable female cooperatives. As one of the existing regencies in Bali Province, Gianyar regency has rapid development of female cooperative. In this research will clustering female cooperative by using WEKA application with Expectation-Maximization Algorithm. From the result, government can maintance which female cooperation should be put on watchlist.**

***Index Terms*—Female Cooperative, Data Mining, Expectation-Maximization, WEKA**

I. INTRODUCTION

Chosen as a demonstration of Project Perkassa (Women Family Health and Prosperous), Gianyar becomes the region with the largest number of cooperatives in Bali Province. There were 77 female coorperatives recorded and registered in the office of coorperatives and SME (Small and Medium Enterprise) Gianyar regency in 2007. Until 2017, there were 147 female cooperatives and there is also a possibility the number increasing.

Cooperative is a collection of people, not a collection of capital [1]. Cooperative must be dedicated for interest of humanity solely and not for material.coorporation in cooperatives is based on sense of equlity and awareness of its members. It is a container of economic and social democrazy. Coorperatives are jointly owned by members, administrators and managers. The business is organized in accordance with the wishes of members through the meeting.

The role of female coorperatives in women's empowerment is the most dominant to provide capital investment credit and surrounding communities in general who want to expand their business or start a business[2]. In addition, female coorperatives can also play an important role in the empowerment of woman among the others, for example provide training, business consultations, improve skill in business technical matters such as organization, management, administration and awareness raising of women on their rights at work environment and family [3]. Because of this importance, the government must be able to categorize the coorperatives so the assistance or counseling can be given on right target.

This research will make a pattern mapping by clustering using Expectation-Maximization (EM) algorithm with WEKA application. With the result, is expected to assist the government in planning new strategies for the advancement of coorperatives.

II. STUDI LITERATURE

A. Literature Review

Many researchs have been done with Expectation-Maximization (EM) algorithm. EM can be used for directing position determination [4] to find find Maximum Likelihood (ML) estimation when the available data can be viewed as incomplete data and when complete data is hidden in the model. Beside it, EM is used too for multi microphone speech dereverberation and noise reduction [5]. EM can do topic modeling on big data streams to measure the centroid for each cluster [6]. EM has a role for pattern recognition in traffic speed distribution which how to understanding driving behavior is a complicated ones [7].

B. Data Mining

The need to extend the capabilities of human analysis to manage the large number of data that human may collect has become increasingly necessary in nowdays [8]. Human ability has a limit to analyse the large databases manually. When the computer invented and allowed for storing more data, the computational techniques for help human was created. How to help human to discover massive structures volume data and the most important is how to understant meaningful patterns. Some limitations, especially to grouped the data, will be appear by taking advantage for grouping the data manually or finished by one man[9].

Data mining is a series of processes to explore the added value of a data set of knowledge that has not been known manually or in other words the process for extracting patterns from the data so as to transform the data into information. It is the process for analysing data from different kind of perspective and make conclusion into useful information for the purpose like how to cut expenses, gain revenue or both[10]. Data mining has many functionalities, such as making data summaries, association analysis among data, data classification, prediction,

(p-issn: 2579-5988, e-issn: 2579-597X)

and data clustering. Each functionality will produce knowledge or patterns that are different from each other.

Data mining has two roles, first is to predict the value of a particular attribute based on the value of other attributes. Predictable attributes are generally known as targets or non-free variables, while the attributes used to make predictions are known as explanatory or independent variables. And the second is descriptive which is decreases patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize key relationships in the data. And is an investigation and often requires postprocessing techniques for validation and explanation of results.

Analysis methods in data mining can be classified by:

1. Association

Search association rules that show the conditions of attribute values that often occur together in a set of data. Associated analyzes are often used to analyze basketball markets and transaction data.

2. Classification and prediction

The process of finding models (functions) that explain and distinguish classes or concepts, with the aim that the model obtained can be used to predict which class or object has an unknown class label.

3. Clustering

Analyze data objects where unknown class labels can be used to determine unknown class labels by way of grouping data to form a new class. Maximizing intra-class similarity and minimizing the interclass resemblance

4. Outliner

Outliers are data objects that do not follow the general behavior of data. Outliers can be regarded as noise or exceptions. Outlier data analysis is called outlier mining. This technique is useful in fraud detection and rare event analysis

5. Trend and evolution

The analysis of data evolution explains and models the trends of objects that have behaviors that change over time. This technique may include characterization, discrimination, association, classification, or clustering of time-related data

C. Clustering

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is the process of collection of abstract or physical objects into a number of classes which are composed of similar objects [11].

Clustering has an important part of data mining, statistical machine learning and pattern recognition [12]. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Final result of data clustering is the unsupervised classification of data into approximately homogenous clusters or groups based on one chosen similarity measurement. With data clustering, key events through process historical databases can be identified and connected to meaningful operating conditions in a plant [13]. To determine a good number of clusters is a challenge that need to be solved and remains as much of an art as a science [14].

D. Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) is known by providing a good benchmark in many machine learning diciplines, for example like image retrieval, speech recognition and natural language processing which is underlying basis of probabilistic inference is their hidden states [4]. EM algorithm can be said as a popular iterative refinement algorithm which can be used to find the parameter estimates. EM begins with an initial estimate of parameters of the mixture model. It will iteratively rescores the objects to against the mixture density created by the parameter vector. Then, the rescored objects are used to update the estimation of parameter. Each object is assigned a probability which would possess a certain set of attribute values by given that it was a member of given cluster. EM algorithms's each iteration consists of two processes. The first is Expectation Step (E-step) and the second is Maximization Step (M-step) [15].

E-step assigns each object x_i to cluster C_k with the probability that can be seen in equation (1). $p(x_i|C_k) = N(m_k,E_k(x_i))$ is following the normal distribution around mean (m_k) with expectation E_k. So this step is calculating the probability of cluster membership of object x_i for each clusters. The probabilitis are the 'expected' cluster membership for object x_i.

$$P(xi \in Ck) = p(Ck|xi) = \frac{p(Ck)p(xi|Ck))}{p(xi)} \qquad (1) \quad [15]$$

M-step uses the probability that estimates from E-step to re-estimate or rescore the model parameters as define in equation (2). This step is the maximization of the likehood of the data distributions given.

$$m_k = \frac{1}{n}\sum_{i=1}^{n} \frac{xiP(xi \in Ck)}{\sum jP(xi \in Cj)} \qquad (2) \quad [15]$$

III. Research Methodology

A. Data Source

In this research, the data used is data achievement performance of Gianyar Regency in accordance with an indicator which has been determined. In Table 1 below is a table of indicator list.

TABLE I
INDICATOR COORPERATIVE

Code	Aspect Indicator
A	District
B	Amount of Female Member
C	Own Equity
D	Outside Equity
E	Business Volume
F	SHU
G	Assets

The data that will use is data result of performance of Female Coorporative in Gianyar in 2016. Then the data will be clustered based on the variable of region and each indicator as seen in Figure 1 . There are 96 cooperatives that will be process

	A	B	C	D	E	F	G
1	kecamatan	anggota_wanita	modal_sendiri	modal_luar	volume_usaha	shu	aset
2	GIANYAR	368	284117000	51834000	317206000	34564000	335951000
3	BLAHBATUH	18	136915645	481700000	106502000	20807645	618615645
4	BLAHBATUH	68	49739000	112487000	71460000	4264000	162226000
5	GIANYAR	19	50194500	4000	19873000	187500	50198500
6	GIANYAR	14	18358000	80375000	89225000	2869000	98733000
7	GIANYAR	18	32682000	4000	12470000	175000	32686000
8	GIANYAR	24	38886000	0	198540000	10332000	38886000

Fig 1. The List of Data that will be Process

B. *Research Steps*

EM algorithm is used in this research, started with study literature looking for supporting and research methods. The next step is how to understanding and what information is needed for analys. After that needed process of understanding and information that needed for analyzed variables. And then the next step is preprocessing data in order to deleting unnecessary data and adding some useful information. After doing preprocessing, the data will be cluster with EM algorithm. The final step is evaluating results of its cluster.

IV. Results and Discussion

A. *Data Preprocessing*

Dominated by increasingly large dataset, data preprocessing and reduction becoming essential techniques. Although being less known than other popular step like data mining, data preprocessing needs more time and effort within the analysis process. Raw data usually have not in perfect conditions, for example they can be inconsistence, value missed, redundant or noise . Transforming raw input to high quality data is the main function of data preprocessing [16]. This is the mandatory step and consist of integration, cleaning, normalization and transformation techniques.

During preprocessing, the availability and characteristics of the data must be determine including the existence of missing value and flat data through the time series [17].

WEKA application can operate the file with arff, c4.5, json and csv format.

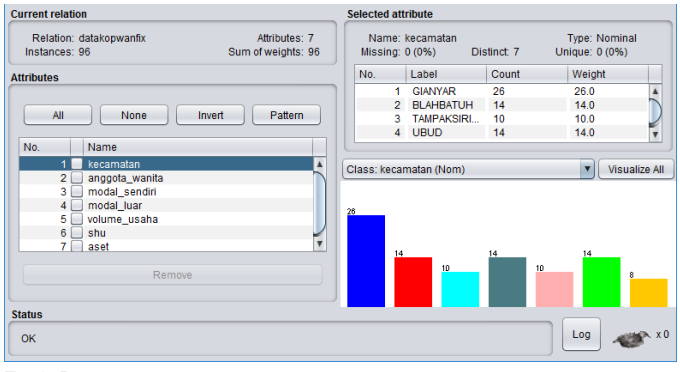

Fig 2. Preprocessing

The first step can be seen in Figure 2, how WEKA application doing the data preprocessing. The data consists of 7 attributes, they are Kecamatan (District), Anggota Wanita (Female Member), Modal Sendiri (Own Equity), Modal Luar (Outside Equity), Volume Usaha (Business Volume), SHU and the last is Aset (Asset).

TABLE II
Weight of District's Attributes

District	Weight
Gianyar	26.0
Blahbatu	14.0
Tampaksiring	10.0
Ubud	14.0
Payangan	10.0
Sukawati	14.0
Tegallalang	8.0

The selected attribute is District as shown in Table II and consist of their weight. The total of the row is 96 rows.

B. *Clustering with EM Method*

After data processing, the EM method of clustering can be choose in WEKA application, with the maximum iteration is setting in 100. The data will be have 3 clusters.

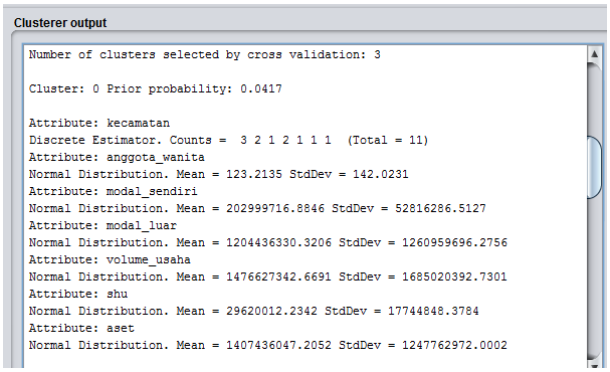

Fig 3. First Cluster

The first cluster (Cluster 0) has prior probability 0.0417. Each attributes contains of the mean and standart deviation as shown in Figure 3.

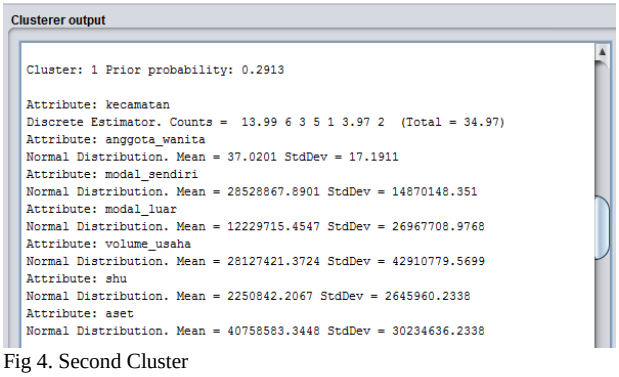

Fig 4. Second Cluster

The second cluster (Cluster 1) has prior probability 0.2913. Each attributes contains of the mean and standart deviation as shown in Figure 4.

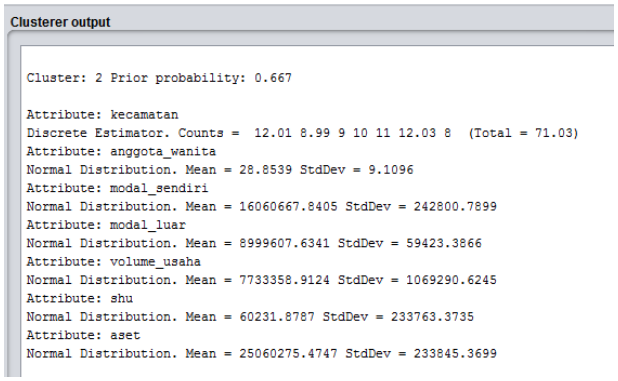

Fig 5. Third Cluster

The third cluster (Cluster 2) has prior probability 0.667. Each attributes contains of the mean and standart deviation as shown in Figure 5.

C. Pattern Mapping

Figure 6 shown how the cluster occurs, it visualization the scaling with cyan until yellow color.

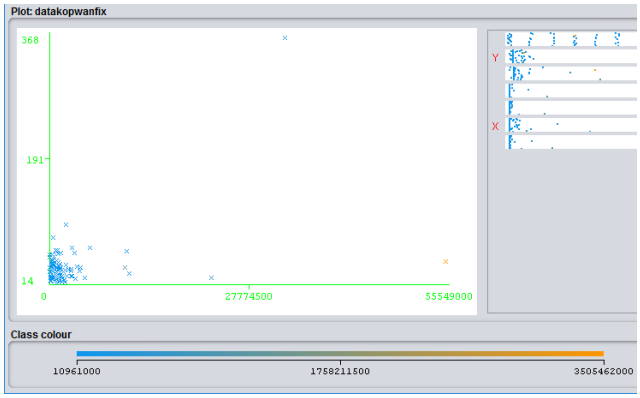

Fig 6. Pattern Mapping

EM is providing the result for clustered instance as shown in Figure 7.

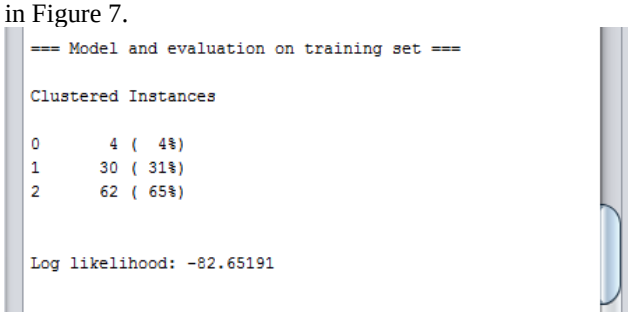

Fig 7. Pattern Mapping

Clusters are represented in visualization, with the clustered instant in first cluster is 4%, second cluster is 31% and third cluster is 65% as shown in Figure 7.

V. Conclusion and Future Works

This research's result with EM algorithm produced 3 cluster area which compotitions are 4%, 31% and 65%. From the analysis, the cooperatives office of Gianyar must have new strategy and policy support the existence and role in female coorperatives. In future research, the combination of another algorithm can make more accurate result in cluster.

References

[1] P. R. Purnaningsih, P. P. Ekonomi, J. P. Ekonomi, and F. Ekonomi, "Profitabilitas Modal Ekuitas Pada Koperasi Wanita As Sakinah Sidoarjo Ekuitas," no. 1997, pp. 1–8, 2013.

[2] F. Meier, "What determines women ' s participation in and within cooperatives ? Evidence from a coffee cooperative in Uganda 1," no. August, pp. 1–34, 2014.

[3] T. Woldu, F. Tadesse, and M. Waller, "Women's Participation in Agricultural Cooperatives in Ethiopia," no. June, p. 22, 2013.

[4] E. Tzoreff and A. J. Weiss, "Expectation-maximization algorithm for direct position determination," *Signal Processing*, vol. 133, no. October 2016, pp. 32–39, 2017.

[5] O. Schwartz, S. Gannot, and E. A. P. Habets, "An Expectation-Maximization Algorithm for Multimicrophone Speech Dereverberation and Noise Reduction with Coherence Matrix Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1491–1506, 2016.

[6] E. J. Moore and T. Bourlai, "Expectation maximization of frequent patterns, a specific, local, pattern-based biclustering algorithm for biological datasets," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 812–824, 2016.

[7] S. C. Lo, "Expectation-maximization based algorithm for pattern recognition in traffic speed distribution," *Math. Comput. Model.*, vol. 58, no. 1–2, pp. 449–456, 2013.

[8] G. Grigoras and F. Scarlatache, "An assessment of the renewable energy potential using a clustering based data mining method. Case study in Romania," *Energy*, vol. 81, pp. 416–429, 2015.

[9] I.C. Dewi, B.Y. Gautama, and P.A. Mertasana, "Analysis of Clustering for Grouping of Productive Industry by K-Medoid Method," *International Journal of Engineering and Emerging Technology*., vol. 2, pp. 26-30, 2017.

[10] I.B.A. Peling, I.N. Arnawa, I.P.A. Arthawan and IGN. Janardana, "Implementation of Data Mining to Predict Period of Students Study Using Naïve Bayes Algorithm ," *International Journal of Engineering and Emerging Technology.*, vol. 2, pp. 53-57, 2017.

[11] C. Jinyin, L. Xiang, Z. Haibing, and B. Xintong, "A novel cluster center fast determination clustering algorithm," *Appl. Soft Comput.*, vol. 57, pp. 539–555, 2017.

[12] K. Zhou, S. Yang, and Z. Shao, "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study," *J. Clean. Prod.*, vol. 141, pp. 900–908, 2017.

[13] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *J. Process Control*, 2016.

[14] P. W. Murray, B. Agard, and M. A. Barajas, "Market segmentation through data mining: A method to extract behaviors from a noisy data set," *Comput. Ind. Eng.*, 2017.

[15] W. Romsaiyud, "Expectation-maximization algorithm for topic modeling on big data streams," *Ubiquitous Comput. Electron. Mob. Commun. Conf. (UEMCON), IEEE Annu.*, pp. 1–7, 2016.

[16] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017.

[17] J. Peral, A. Maté, and M. Marco, "Application of Data Mining techniques to identify relevant Key Performance Indicators," *Comput. Stand. Interfaces*, 2016.