International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 38314-38319
© Research India Publications. http://www.ripublication.com

Optimization of K-Prototype Clusters with Genetic Algorithm for Customer Segmentation

I Made Ari Santosa
Lecturer,
Department of Computer System
STIKOM Bali College of Information Management and Computer Engineering
Bali, Indonesia
arisantosa@stikom-bali.ac.id

Abstract— Clustering is a data mining techniques that widely used by companies to group and map their market heterogeneity and customer needs. Many solutions have been proposed to deal with problems associated with clustering market heterogeneity and customer needs. In this paper the use of genetic algorithms based k-prototype clustering is proposed with the aim to get better clusters quality and optimize the number of clusters. This approach consists of solution to optimize data cluster which is expected to produce the optimal number of clusters in order to provide better data segmentation results.

Keywords: data mining, customer segmentation, clustering, k-prototype, genetic algorithm, optimization.

Introduction

Customers are very important element for the development of the organization or company, because the customers are a source of revenue and profits for the organization or the company itself [1]. Information and knowledge of the customer's needs that increasingly diverse is indispensable, especially in the groups or segments of potential and profitable customers for the company [2],[3]. Information and knowledge about the customers is very useful for the company to plan appropriate business strategy [3]. Therefore, customers or market segmentation has become a very fundamental issue in line with the increasingly diverse of customer behavior and needs [4],[5].

The strategy which is commonly used by companies or marketers to group and map their market heterogeneity and customer needs is segmentation [4],[6]. Segmentation means dividing the overall customers or market into some small groups with a degree of similarity of characteristics or a certain behavior [2],[4]. With the customer segmentation, businessman and companies can more easily to plan a marketing strategy based on the character or behavior of any existing group or segment [2]. Customer segmentation will also greatly assist the company in providing a supply of goods or services to the right customer [3].

One of the most popular data mining techniques that can be use for market or customer segmentation is cluster analysis [7],[8]. Cluster analysis is a technique that is often used to divide or classify data into groups or segments based on characteristics or with certain criteria [4]. Cluster analysis is used to divide a set of objects or data into some of particular groups, so that the values in each group is homogeneous, which refers to certain attributes based on certain criteria [7],[9]. The same value or similarities that found in each segment or cluster describe similarities and behavioral patterns of customer transactions [7],[9]. The accuracy of each group or cluster which generated shows the similarity degree of the patterns of customer behavior [10].

There are various algorithms which can be used to perform grouping or clustering of data, one of which is the k-means clustering, which is the most popular clustering algorithms and has its application in data mining, image segmentation, and bioinformatics as well [1],[4],[11]. The advantages of k-means is easy to implement because the algorithm is simple and efficient enough to solve clustering problems [1],[11],[12]. However, although the K-means clustering (and also other cluster algorithms) have been used widely, but the problems associated with variable and also produced the optimal number of clusters is not widely discussed in the analysis of customer segmentation as well as in the analysis of market segmentation [13].

To overcome the problems associated with the optimum cluster number which can be generated, k-means clustering method combined with genetic algorithm is proposed in [13]. This approach aim to be a solution to obtain the optimum cluster number simultaneously [13]. However, the proposed method is not considered the possibility of distortion of data, as a result of differences in data types used in the data set. It is because not all categorical data in the real world is presented in the order form, and categorical data handling by methods for numerical data often gives result which is not optimum [14],[15],[16]. To that end, a clustering algorithm called the k-prototype proposed to deal with clustering on the mixed data type which is numerical and categorical [14],[15].

In this study, the use of genetic algorithms in k-prototype clustering is proposed with the aim to optimize the number of clusters generated by the k-prototype algorithm. This approach consists of solution to optimize data cluster which is expected to produce the optimal number of clusters in order to provide better data segmentation results.

Related Work

Clustering is one of method for grouping data which is known as unsupervised classification, because there is no training phase necessary for this method. Unsupervised classification is a classification method that does not require labeling class of each data to be investigated [2]. There are two approaches in clustering method, those are, the clustering method based on partitioning and clustering methods based on hierarchy [12]. Some related works to address the problems deals with cluster optimization for data segmentation and also associated with the use of mixed data type been done.

Widiarini et al, as in paper [1], proposed a dynamic cluster algorithm in k-means algorithm, to solve the problem of the sensitivity of the initial partition number of clusters on the k-means algorithm. The use of dynamic cluster algorithm for optimization of k-means cluster was able to produce better cluster quality. In principle, the proposed method is using the k-means algorithm to obtain the desired cluster and then the calculations of intra and inter clusters to get a better cluster quality [1],[17]. A study has conducted, as in [9], used the combination of similarity weight based clustering algorithm and filter method paradigm to perform clustering on the data with mixed attributes. Methods that they used in this study are divide the data sets with mixed attributes into 2 (two) parts, those are, data sets with numeric attributes and data sets with categorical attributes. And then, they do the clustering on each of data sets by using Similarity Weight algorithms. Clusters which they obtained, (from the data sets with numeric and categorical attributes) will be combined and then they performed clustering using the filter method. The results of the experiment show that algorithm has been used generates better cluster quality than k-prototype algorithms [9]. A similar study also conducted by Ahmad et al, as in [18], using the k-means algorithm to perform clustering on the data set with a mix of numeric and categorical attributes. Research that has been conducted using different approaches on measurement of the distance cluster that can be used in the data set with a mix of numeric and categorical attributes. To overcome the weaknesses of k-means algorithm which is limited to numerical data sets, they have made modifications to the cluster center [18]. A modification on the cluster center also produces better clusters characteristic. Another related work deal with clustering mixed data set also conducted as in [19] by using a combination of genetic algorithm and k-mean. In this research, Roy et al, as in paper [19], have made a modification on the description of the cluster center as practiced in [18] to overcome the weaknesses of k-means algorithm which is limited to the data sets with numerical attributes. Unlike the previous studies conducted as in [18], on this research, they utilize genetic algorithms to optimize cluster on a data sets with a large amount [19].

A related studies deal with market segmentation was proposed in paper [3] by using a combination of k-means algorithm based fitness function, Particle Swarm Optimization (PSO). This study aims to provide an effective and accurate market segmentation system.

In addition, there is also research conducted by Kuo et al, as in [20], by using two methods: Self-Organizing Feature Maps (SOM) neural network to determine the starting point and the number of clusters, and then use a combination of genetic algorithm and k-means (GA K-means) to find a final solution. Based on the research that has been conducted shows the combined method GA K-means segmentation showed better results compared to methods based on neural network. Another similar study also conducted by S. Balaji et al, as in [2], using the association rules approach. This research aims to perform segmentation and customer behavior predictions in order to develop the right products for customers. The study was conducted through three stages, namely, the data preparation, clustering of data and customer preference analysis.

Proposed Method

K-prototype cluster based on genetic algorithm, as the proposed method in this study, refers to the approach by Liu, as in paper [13], with changes on the clustering algorithm (k-mean) and fitness evaluation. In this study we used k-prototype as proposed in [15], as a clustering algorithm and using a cost function as proposed in paper [14].

K-prototype cluster based on genetic algorithm optimization method begins with initialization parameter, in which the chromosomes are divided into two parts. The number of genes contained in the chromosomes is equal to the number of attributes or variables of the data set that will be used. This gene is represented by the numbers of 1 to indicate that the attribute which represented will be included in the process of clustering, while the number of 0 indicates that the attribute which represented will not included in the clustering process [13]. Furthermore, random population initialization is performed. Population initialization phase is used to determine the beginning number of chromosomes to be used for the next computing. Initial determination of chromosome is done randomly [10],[13],[21].

Pre-processing process is performed on the data sets before the clustering process. Processes performed at this stage is to eliminate the class label, normalization of the numerical data type, representation change the of categorical data type into numeric [10],[22]. The next stage is performed clustering process using the k-prototype algorithm.

K-prototype algorithm is one of the clustering methods based on partitioning. This algorithm is a result of enhancement and integration of k-means and k-modes clustering to handle the data sets with mixed type of attributes (numeric and categorical attributes). There is a fundamental change in the measurement of similarity between the objects with the centroids (prototype) [14],[15].

K-prototype algorithm consists of several steps namely, prototype initialization of the data set, allocation of the objects in X to the cluster with the closest prototype, perform object reallocation if there is a prototype change, if the center point of the cluster has not changed or has been convergent, then the algorithm stops, but if the center point of the cluster is still changing significantly the process back to the previous stage until there is no longer an object displacement or changes in cluster [15]. Fig.1. shows the steps of k-prototype algorithm.

At this stage of clustering, each chromosome that is formed in the initialization phase of population is going through the process of clustering. Clustering Criterion will be used as a fitness value of each chromosome that has been evaluated. In this study, the clustering criterion which used is the cost function, or the cost which spent on placing objects in the corresponding cluster. Huang cost function is used as a formula for calculating the cost function as in [14]. The formula used is as follows:

$$P(w,Q)=\sum_{l=1}^{k}\left(\sum_{i=1}^{n}y_{il}\sum_{j=1}^{m_r}(x_{ij}^{r}-q_{lj}^{r})^{r}+\gamma_l\sum_{i=1}^{n}y_{il}\sum_{j=1}^{m_c}\delta(x_{ij}^{c},q_{lj}^{c})\right) \quad (1)$$

Let $$\sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} (x_{ij}^{r} - q_{lj}^{r})^{r} = E_i^r$$

And $$\gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^{c}, q_{lj}^{c}) = E_i^c$$

And $$P(w,Q) = Cost_{\,i}$$

Then the equation (1) can be written as follows:

$$Cost_{\,i} = \sum_{i=1}^{k} (E_i^r + E_i^c) \qquad (2)$$

Where E_i^r is the total cost for all the numerical attributes of the objects in the cluster i and E_i^c is the total cost for all the categorical attributes of the objects in the cluster *i*. Thereby, the equation (2) can be written as follows:

$$Cost_{\,i} = \sum_{i=1}^{k} E_i^r + \sum_{i=1}^{k} E_i^c \qquad (3)$$

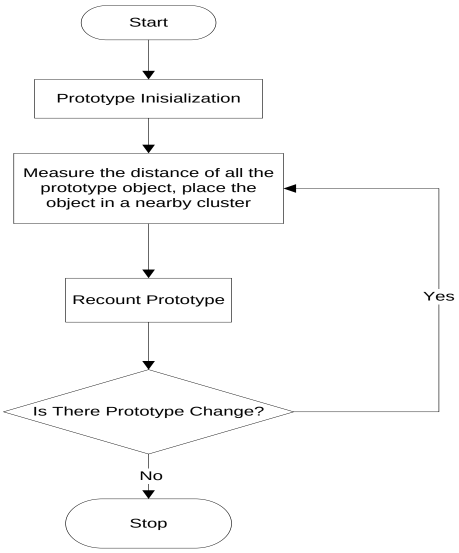
Start
Prototype Inisialization
Measure the distance of all the prototype object, place the object in a nearby cluster
Recount Prototype
Is There Prototype Change?
Yes
No
Stop

Fig.1. K-prototype algorithm [15]

In order to obtain chromosomes or individual who has the best fitness value, then the iteration process of reproduction is performed. The iteration process performed as much as the number of generations that have been determined. To determine the chromosome to be used as a parent in the crossovers process, then the selection process is performed [13]. Wheel Roulette used as model selection in this study, because this model provides a greater probability to the chromosomes that have higher fitness value [21].

In the genetic algorithm, crossover operator has the most important role, because in crossover operator there is the process of mating (crossing) genes between two individuals (the parent) which resulted in two new individuals (offspring) in the next generation [13],[23]. Mutation process performed to create a new individual; thereby the population variation will be increasing. Mutation process is conducted by random modification of one or more genes in the same individual. Mutation is useful to replace the missing genes from the population during the selection process, as well as providing a gene that does not exist in the initial population [13].

Experiment Results

The experiment has been conducted to evaluate the proposed method in large data sets. In this study, the data sets used is German Credit Data Set. This data set can be downloaded in the UCI Machine Learning Repository [24]. German Credit Data Set is the data sets donated by Prof. Hofman from Hamburg University, Germany. The data set consists of 1000 records with 20 attributes which 13 categorical attributes and the remaining 7 attributes are numeric [13],[24].

Results of experiments performed, will determine the optimal number of clusters, and the highest fitness value of each experiment as well as the best form of chromosomes that will be generated. Tests performed on the parameters of population size and maximum of generation. Tests on the maximum parameter generation made with different values ranging from 50 to 1000 generation. From the test results in the maximum parameter generation in Table.2, it can be analyzed that there is a trend of increasing fitness value generated along with the maximum number of generations.

TABLE.2. Changes of fitness value in each generation

MAXIMUM OF GENERATION	FITNESS VALUE
50	0.00081868
100	0.00085543
200	0.00096376
300	0.00092556
400	0.00089412
1000	0.00099521

Experimenting with population size parameter also conducted to see changes in the value of fitness based on the size of the population being used. From the test results, as shown in table.3, obtained the best fitness value of 0.00108870 in the parameter size of the population in 1000.

TABLE.3. Population size and changes of fitness value

POPULATION SIZE	FITNESS VALUE
50	0.00081868
100	0.00090613
200	0.00087169
400	0.00092556
500	0.00104140
1000	0.00108870

The test results showed that the optimum fitness value obtained 0.00108870 with the best chromosome 01001101011011111111 value as shown in Fig.2.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1	1	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1

Fig.2. Best Chromosome

The gene that is represented by the binary number 1 denotes as selected attributes to be included in the process of clustering, whereas binary 0 indicates as attributes which not included in the clustering process and considered as not significant variables [13]. Based on the value of the best chromosome obtained, then the attributes (variables) that are not included in the clustering process are: status of existing checking account, purpose, savings account/bonds, present employment since, personal status and sex, property.

Clustering validity index is used to measure the quality of the generated clusters. Method used to measure the validity of the generated clusters are Category Variance Criterion proposed by Hsu et al, as in [8], which is a combination between Category Utility methods and variance measurement for numerical data[8],[25]. The formulas used by Hsu et al, as in paper [8], to measure the validity of the cluster are as follows:

$$CV = \frac{CU}{1+Variance} \tag{4}$$

The purpose of the categorical utility (CU) function is to maximize the likelihood that the two objects in the same cluster have the same attribute values, and two objects in different clusters have different attributes. Categorical utility for a dataset can be calculated as follows [8]:

$$CU = \sum_k \left(\frac{|C_k|}{|D|} \sum_i \sum_j \left[P(A_i = V_{ij} \mid C_k)^2 - P(A_i = V_{ij})^2 \right] \right) \tag{5}$$

Where $P(A_i = Vij \mid C_k)$ is conditional probability where the attribute i has the value V_{ij} in the cluster C_k, and $P(A_i = V_{ij})$ is the probability of overall that the attributes i has the value V_{ij} in the entire data set. This function aims to measure if clustering increases the likelihood of the same value is in the same cluster. Thereby, along with the higher value of the CU, it will get better, as well as the quality of the clustering [8].

From the results of experiments that have been conducted, k-prototype which optimized by a genetic algorithm generated better cluster quality than the k-means based on genetics algorithm, with a Variance Category value of 0.24387. Table.4 shows that the k-prototype-based genetic algorithm has a higher Category Variance Criterion value than the k-means based on genetics algorithm. In addition, the total cost required is also less when compared with the k-means based on genetics algorithm, as shown in table.5. The number of clusters generated at k-prototype based genetic algorithms as many as 8 clusters or segments. A comparison the number of clusters and the number of elements from each cluster can be seen in table 6, and also tabulated in the form of a chart as shown in Fig.3.

TABLE.4. Cluster Numbers and CV Index

ALGORITHM	CLUSTER NUMBERS	CV INDEX
GA K-prototype	8	0.24387
GA K-means	4	0.21853

TABLE.5. Best Chromosome and Total Cost

ALGORITHM	BEST CHROMOSOME	TOTAL COST
GA K-prototype	01101001011011111111	918.5462
GA K-means	11111011110110101101	2065.3745

TABLE.6. Number of elements from each cluster

ALGORITHM	CLUSTER NUMBERS	CLUSTER	TOTAL OF ELEMENT
GA K-Prototype	8	1	43
		2	123
		3	84
		4	80
		5	218
		6	269
		7	65
		8	118
GA K-Means	4	1	153
		2	279
		3	36
		4	532

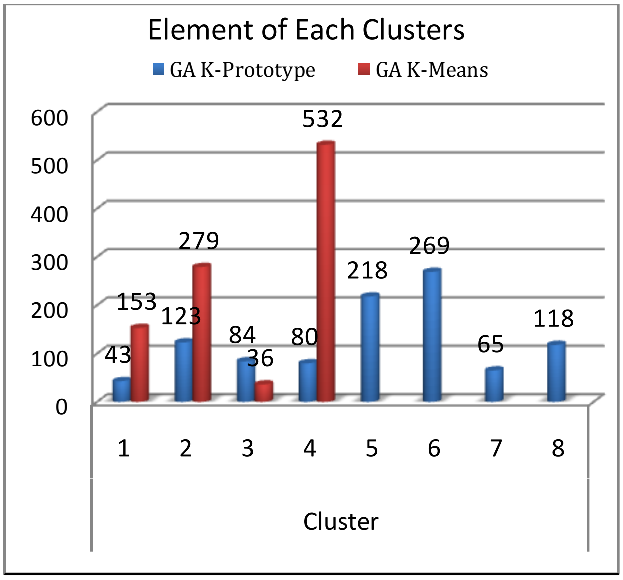

Fig.3. Comparison Chart of cluster numbers and the elements of each cluster

Based on the data sets that have been used, description from each cluster (segment) which generated at k-prototype-based genetic algorithm can be described as follows:

a. Segment 1, the number of customers are as many as 43 dominated by foreign workers, has no people being liable to, self-employed and no registered phone number.
b. Segment 2, the number of customers are as many as 123 is dominated by customer which has no people being liable to, some of them are foreign workers, some of them have no other installment plans, housing status is rent, and no other debtors / guarantors.
c. Segment 3, the number of customers are as many as 84 dominated by foreign workers customer, other debtors/guarantors are none, has no people being liable to, some of them are self employed and highly qualified employee/officer with the telephone registered under the customer's name.
d. Segments 4, the numbers of customers are as many as 80 dominated by foreign workers which some of them have number of people being liable to, some others are skilled employee/official, and other debtors/guarantors are none.
e. Segment 5, the number of customers are as many as 218 is dominated by foreign workers which some of them have number of people being liable to, other debtors / guarantors is none, most of them have no other installment plans, most customers have their own house and phone number registered under the customer's name.
f. Segment 6, the number of customers are as many as 269 is dominated by customers which has no number of people being liable to, some of them are foreign workers, some customers already have their own home, some of them do not have another guarantors and do not have another installment plans.
g. Segment 7, the number of customers are as many as 65 is dominated by customer which has no number of people being liable to, most of them are foreign workers and has no other debtors / guarantors, some customers already have their own home and do not have another installment plans.
h. Segment 8, the number of customers are as many as 118 is dominated by customer which has no number of people being liable to, some of them are unskilled – resident, some others are foreign workers, and some others are customer which already have their own home and do not have another installment plans.

Conclusion

From the experiment that has been conducted show that the k-prototype with genetic algorithms could perform clustering on mixed attributes data set with optimal result. Based on the results of experiments that have been done, k-prototype with genetic algorithms could generate clusters with better accuracy. The quality of the generated clusters is based on the value of category variance criterion which is obtained. In addition, based on the total cost result, k-prototype with genetic algorithms performs more effective clustering process for data sets with mixed attributes. In general, for data sets with mixed attributes, k-prototype with a genetic algorithm generate a more optimal number of clusters compared with the k-means based on genetic algorithm.

Acknowledgment

Thanks to Prof. Richardus Eko Indrajit, from ABFI Institute of Perbanas, Dr. Bambang Eka Purnama, from Surakarta University, and I Wayan Budi Sentana, M.Sc., from Bali State Polytechnic.

References

[1] Widiarini and R. Satria Wahonono, "Algoritma Cluster Dinamik Untuk Optimasi Cluster Pada Algoritma K-Means Dalam Pemetaan Nasabah Potensial," *Journal of Intelligent Systems*, vol. 1, no. 1, pp. 32–35, 2015.

[2] S. Balaji and S. K. Srivatsa, "Customer Segmentation for Decision Support using Clustering and Association Rule based approaches," *International Journal of Computer Science and Engineering Technology*, vol. 3, no. 11, pp. 525–529, 2012.

[3] C.-Y. Chiu, Y.-F. Chen, I.-T. Kuo, and H. C. Ku, "An Intelligent Market Segmentation System Using K-Means and Particle Swarm Optimization," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4558–4565, 2009.

[4] R. J. Kuo, L. M. Ho, and C. M. Hu, "Integration of Self-Organizing Feature Map and K-Means Algorithm for Market Segmentation," *Computers & Operations Research*, vol. 29, no. 11, pp. 1475–1493, 2002.

[5] M. Y. Kiang, M. Y. Hu, and D. M. Fisher, "An Extended Self-Organizing Map Network for Market Segmentation-a Telecommunication Example," *Decision Support Systems*, vol. 42, no. 1, pp. 36–47, 2006.

[6] M. Tuma and R. Decker, "Finite Mixture Models in Market Segmentation: A Review and Suggestions for Best Practices," *Electronic Journal of Business Research Methods*, vol. 11, no. 1, pp. 2–15, 2013.

[7] J. J. Huang, G. H. Tzeng, and C. S. Ong, "Marketing segmentation using support vector clustering," *Expert Systems with Applications*, vol. 32, no. 2, pp. 313–317, 2007.

[8] C. C. Hsu and Y. C. Chen, "Mining of Mixed Data with Application to Catalog Marketing," *Expert Systems with Applications*, vol. 32, no. 1, pp. 12–23, 2007.

[9] M. V. J. Reddy and B. Kavitha, "Clustering the Mixed Numerical and Categorical Datasets Using Similarity Weight and Filter Method," *International Journal of Database Theory and Application*, vol. 5, no. 1, pp. 121–133, 2012.

[10] I. M. A. Santosa and I. W. B. Sentana, "Kombinasi Algoritma Genetik dan K-Prototype untuk Menentukan Jumlah Cluster Optimal pada Data Bertipe Campuran," *Koferensi Nasional Sistem Informasi*, pp. 1354–1363, 2014.

[11] S. S. Raghuwanshi, P. Arya, M. T. Ss, C. Application, and M. P. Vidisha, "Comparison of K-means and Modified K-mean algorithms for Large Data-set Abstract□ :,"*International Journal of Computing, Communications and Networking*, vol. 1, no. 3, pp. 106–110, 2012.

[12] G. Komarasamy and A. Wahi, "A New Algorithm for Selection of Better K Value Using Modified Hill Climbing in K-Means Algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 55, no. 3, pp. 307–314, 2013.

[13] H. Liu and C. Ong, "Variable Selection in Clustering for Marketing Segmentation Using Genetic Algorithms," *Expert Systems with Applications*, vol. 34, pp. 502–510, 2008.

[14] Z. Huang, "Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998.

[15] Z. Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values," *Proceeding of the First Pacific Asia Knowledge Discovery and Data Mining Conference, World Scientific, Singapore*, pp. 21–34, 1997.

[16] J. Lim, J. Jun, S. H. Kim, and D. Mcleod, "A Framework for Clustering Mixed Attribute Type Datasets," in *Proceeding of the Fourth International Conference on Emerging Database*, 2012.

[17] A. S. B. M and K. S. Hareesha, "Dynamic Clustering of Data with Modified K-Means Algorithm," in *International Conference on Information and Computer Networks*, 2012, vol. 27, pp. 221–225.

[18] A. Ahmad and Dey Lipika, "A K -Mean Clustering Algorithm for Mixed Numeric and Categorical Data," *Data & Knowledge Engineering*, vol. 63, pp. 503–527, 2007.

[19] D. K Roy and L. K Sharma, "Genetic K-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets," *International Journal of Artificial Intelligence & Applications*, vol. 1, no. 2, pp. 23–28, 2010.

[20] R. J. Kuo, Y. L. An, H. S. Wang, and W. J. Chung, "Integration of Self-Organizing Feature Maps Neural Network and Genetic K-Means Algorithm for Market Segmentation," *Expert Systems with Applications*, vol. 30, no. 2, pp. 313–324, 2006.

[21] F. Pasila, D. Gunawan, and H. Ferdinando, "Evolutionary Algorithm pada Fuzzy Clustering Systems Metode Gustafson-Kessel untuk Forecasting Electrical Load Data Time-Series," *Industrial Electronics Seminar*, pp. 15–20, 2008.

[22] F. K. Wardhani, E. Suryani, and A. Mukhlason, "Penerapan metode GA-Kmeans untuk Pengelompokan Pengguna pada Bapersip Provinsi Jawa Timur," *Jurnal Teknik ITS*, vol. 1, pp. 545–550, 2012.

[23] S. Sharma and S. Rai, "Genetic K-Means Algorithm – Implementation and Analysis," *International Journal of Recent Technology and Engineering*, vol. 1, no. 2, pp. 117–120, 2012.

[24] German Credit Data Set,
http://www.cse.ust.hk/~qyang/221/Assignments/German/

[25] C. Hsu and Y. Huang, "Incremental Clustering of Mixed Data Based on Distance Hierarchy," *Expert Systems with Applications*, vol. 35, pp. 1177–1185, 2008.