Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 677–684

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering

Rani Nooraeni[a,*], Muhamad Iqbal Arsa[b], Nucke Widowati Kusumo Projo[a]

[a]STIS Polytechnic Statistic, Jakarta, Indonesia
[b]BPS Statistic Indonesia , Indonesia

Abstract

Statistical data analysis in machine learning and data mining usually uses the clustering technique. However, data with both attributes or mixed data exists universally in real life. K-prototype is a well-known algorithm for clustering mixed data because of its effectiveness in handling large data. However, practically, k-prototype has two main weaknesses, the use of mode as a cluster center for categorical attributes cannot accurately represent the objects, and the algorithm may stop at the local optimum solution because affected by random initial cluster prototypes. To overcome the first weakness, we can use fuzzy centroid, and for second weakness is to implement the genetic algorithm to search the global optimum solution. Our research combines the genetic algorithm and Fuzzy K-Prototype to accommodate these two weaknesses. We set up two multivariate data with high correlation and low correlation to see the robustness of the proposed algorithm. According to four value indexes of clustering result evaluation, Coefficient Varians Index, Partition Coefficient, Partition Entropy, and Purity, show that our proposed algorithm has a better result than K prototype. Based on the evaluation result, we conclude that our proposed algorithm can solve two weaknesses of the k-prototype algorithm.

Keywords: Data Mining; Fuzzy K Prototype; Genetic Algorithm; Mixed Data; Clustering

* Corresponding author.
E-mail address: raninoor@stis.ac.id

10.1016/j.procs.2021.01.055

1. Introduction

Data mining is a multidisciplinary field with a broad range from statistics to data visualization [1]. It also includes database technology, artificial intelligence, pattern recognition, machine learning, information theory, knowledge acquisition, information retrieval, high-performance computing. Data mining aims to extract implicit, previously unknown, and possibly beneficial patterns in data.

Clustering is an essential technique in data mining. It intends to divide objects into groups. The objects are grouped by considering the intra-class similarity and inter-class similarity [2]. Objects in one cluster should have high similarity between each other while they should relatively dissimilar with objects in different groups. Many fields, such as image analysis [3], bioinformatics [4], machine learning, data mining, and pattern recognition [5], have applied clustering techniques for analyzing statistical data.

Recent researches related clustering have been done. Several techniques have been developed, but the existing clustering algorithms mostly focus only on a single data type, either numeric or categorical type. However, the mixed data type is more common in the real world. Therefore, clustering algorithms to handle mixed data is needed.

With the emergence of extensive data that contains mixed attributes, data mining researchers have been trying to find proper criteria function to deal with this problem [6]. Various strategies have been applied, such as converting categorical and nominal attributes into numeric integer values. The similarity between two objects is found by using numeric distance. However, this strategy has an obstacle in precisely assigning numerical values to categorical values such as marital status and gender. A different approach is by discretizing numeric attributes and applying categorical clustering algorithms. However, this process might result in information loss.

The SBAC (Similarity-Based Agglomerative Clustering) algorithm is built upon the Goodall similarity measure [7], which could manage mixed data. However, its quadratic computational cost is expensive when it deals with large data [8]. In 1998, Huang proposed a cost function to handle mixed data. His cost function is an integration of Euclidean distance and simple matching dissimilarity. Then, Huang introduced a clustering method called K-prototype to implement his cost function. By adopting the partition clustering techniques, his algorithm could manage extensive data.

However, Huang's cost function has several shortcomings. First, the mode represents the cluster center of the categorical attribute. Since the determination of the center depends only on the mode, the center cannot accurately portray the objects [6]. Also, when there is an attribute that has more than one mode, the mode then is not unique. Choosing different mode values to be center can give different results, which can lead to an unstable algorithm. Measuring the distance between two categorical objects using binary value cannot accurately describe the similarity of objects. In addition to the previous issues, the second and the most common problem with partition clustering algorithms is they produce the local optimum solution, which is affected by initial cluster prototypes [9]. Besides that, the K Prototype is famous for the dataset with high dimension big observation and large attribute. Nevertheless, it has high risks of multicollinearity or a high correlation between attributes. In some literature, this condition will decrease the clustering performance.

According to the problems described above, it is necessary to develop algorithms based on K-Prototype to handle mixed data. Our study uses fuzzy centroid [10] for categorical attributes to overcome the first weakness. Ji et al [11] implement a fuzzy centroid to their algorithm (Fuzzy K-Prototype). Therefore, we will use their algorithm in this study.

Nooraeni combines the genetic algorithm and k-prototype algorithm. The experiment shows that a genetic algorithm can improve the k-prototype algorithm [12]. For the second weakness, we propose using the genetic algorithm to optimize the clustering algorithm so that it can reach the global optimum [13]. Moreover, we combine Fuzzy K-Prototype and Genetic Algorithm to accommodate those two weaknesses. Lastly, we will train our proposed algorithm to the data with condition high and low correlation. We also will implement the algorithm to the real-world data to test the performance of our algorithm.

2. Methodology

2.1. Similarity Measures

For mixed data types, the popular distance measure is the K-Prototype proposed by Huang [9]. Let $X = \{X_1, X_2, \dots, X_n\}$ denotes a set of n objects and $X_i = [X_{i1}, X_{i2}, \dots, X_{im}]$ is an object represented by m attribute values, then Huang cost function is:

$$d(X_i, Q_l) = \sum_{j=1}^{p}(x_{ij}^{r} - q_{lj}^{r})^2 + \gamma_l \sum_{j=p+1}^{m} \delta(x_{ij}^{c}\ ,\ q_{lj}^{c}) \tag{1}$$

The x_{ij}^{r} are values of numeric attributes and $x_{ij}^{c}$are values of categorical attributes. The cluster center for cluster l is represented with $Q_l = (q_{l1}, q_{l2}, \dots, q_{lm})$.The mean of numeric attributes j and *cluster* l is represented with q_{lj}^{r}. While the q_{lj}^{c} is the mode of attributes j and *cluster* l. The $\delta(p, q) = 0$ for $p = q$,, and $\delta(p, q) = 1$ for $p \neq q$, for categorical attributes. The weight for categorical attributes of cluster l is γ_l.

2.2. Fuzzy clustering and fuzzy centroid

Data objects can be part of many clusters in fuzzy clustering. Each data object has a membership degree that implies a strong relationship between the data object and a specific cluster. Minimizing the objective function is the purpose of the fuzzy clustering algorithm, and it is written as follows:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} {v_{il}}^{\alpha} d(X_i\ , Q_l) \tag{2}$$

where v_{il} defines the grade of membership i data object to the cluster l. α is the fuzziness coefficient.

In hard centroid, each attribute of the centroid has a single hard category value. Contrary to the hard centroid, it has a fuzzy category value in each attribute to characterize the cluster's distribution information. For Dom (A_j) = $\{a_j^1, a_j^2, \dots\ , a_j^t\}$, the $\tilde{\boldsymbol{V}}$ fuzzy centroid can be written as:

$$\tilde{\boldsymbol{V}} = [\tilde{v}_1, \dots, \tilde{v}_j, \dots, \tilde{v}_m] \tag{3}$$

where

$$\tilde{v}_j = a_j^1/\omega_j^1 + a_j^2/\omega_j^2 + \dots + a_j^K/\omega_j^K + \dots + a_j^t/\ \omega_j^t \tag{4}$$

subject to

$$0 \leq \omega_j^K \leq 1,\ \ 1 \leq K \leq t, \tag{5}$$

$$\sum_{k=1}^{t} \omega_j^K = 1,\ 1 \leq j \leq m \tag{6}$$

Zadeh [14] proposes a convenient notation for a fuzzy $\{a_j^K, \omega_j^K\}$, that represents the fuzzy category value of $\tilde{v}_j \in \tilde{V}$. The ω_j^K is a confidence degree measuring how much the contribution of a_j^K to $\tilde{v}_j$.

2.3. Distance and significance

When a data set has m attributes where both categorical and numerical attributes have been discretized, the distance between two distinct values x and y from any categorical attribute A_i is:

$$\delta(x, y) = \left(\frac{1}{m-1}\right) \sum_{j=1, i \neq j}^{m} \delta^{ij}(x, y). \tag{7}$$

The properties for $\delta(x,y)$ are $0 \leq \delta(x,y) \leq 1$, $\delta(x,y) = \delta(y,x)$, and $\delta(x,x) = 0$. The $\delta^{ij}(x,y)$ is the distance between attribute values x and y of A_i as a function of their co-occurrence probabilities, with a set of values of another categorical attribute A_j. We can rewrite as follows:

$$\delta^{ij}(x,y) = p_i(\vartheta/x) + p_i(\sim\vartheta/y) - 1.0. \tag{8}$$

The significance in this algorithm will be applied to the distance function for numeric attributes $d(x,y) = (w_i(x-y))^2$. To calculate the significance attribute, we first discretize the numeric attribute. The discretizing is for calculating the significance attribute and not for the clustering purpose. Because of discretizing, the same number of intervals T *is* chosen for all numeric attributes. A categorical value of $u[1], u[2], \dots, u[T]$ defines each range. We will calculate $\delta(u[r], u[s])$ for every pair of categorical values $u[r], u[s]$. Calculating the discretized numeric attribute uses the same way as categorical values, which is mentioned in Equation (7). The last step is computing the significance of a numeric attribute by averaging of all pairs $\delta(u[r], u[s])$ as:

$$w_i = \frac{2\sum_{k=1}^{T}\sum_{j>k}^{T}\delta(u[k],u[j])}{T(T-1)} \tag{9}$$

2.4. Fuzzy k-prototype

The first weakness of K-prototype is using the mode to represents the cluster center for categorical attributes. To solve this problem, we propose using a fuzzy centroid. Ji et al [11] implement fuzzy centroid to their algorithm (Fuzzy K-Prototype Algorithm). The objective function can be rewritten as:

$$E(U,Q) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{\alpha}\left(\sum_{l=1}^{p}(w_l(x_{il}^{r} - q_{jl}^{r}))^2 + \sum_{l=p+1}^{m}\varphi(x_{il}^{c}, \tilde{v}_{jl}^{c})^2\right) \tag{10}$$

The calculation of w_l is using Equation (12). q_{jl}^{r} is numeric attribute of the centroid of, v_{jl}^{c} is the categorical attribute of the fuzzy centroid, , where $\delta(x_{il}^{c}, a_{jl}^{K})$ is calculated by Equation (8).

2.5. Optimize Solutions with Genetic Algorithms.

We propose the genetic algorithm to optimize the results of clustering with Fuzzy K Prototype so that the resulting algorithm is a hybrid algorithm named Genetic Fuzzy K Prototype Algorithm (GAFKP algorithm). GAFKP consists of seven elements: *coding representation*, *population initialization*, *evaluation of fitness value*, *selection*, *crossover*, *mutation*, and *elitism*. We use a one-step Fuzzy K-Prototype algorithm as the crossover operator to accelerate the convergence process. The elements process is described following the work of Zhao et al [15]:

- *Coding representation*
  The $\boldsymbol{k} \times \boldsymbol{n}$ fuzzy membership matrix $\boldsymbol{W}$ represents a chromosome, where $\boldsymbol{k}$ is the cluster's number, and $\boldsymbol{n}$ is the object's number in the dataset. When we have $\boldsymbol{k} = \mathbf{3}$ and $\boldsymbol{n} = \mathbf{4}$, it means that the chromosome $(\boldsymbol{a_{11}}, \boldsymbol{a_{21}}, \dots, \boldsymbol{a_{kn}})$ is the $\mathbf{3} \times \mathbf{4}$ fuzzy membership matrix.
- *Initialization process*
  To produce the initial chromosome $(a_1, a_2, \dots, a_{n.k})$ where n is the population size. Produce k random numbers $v_{i1}, v_{i2}, \dots, v_{ik}$ from [0,1]. Then For the i th point of the chromosome. Compute $a_{(j-1)*n+i} = v_{ij}/\sum_{l=1}^{k} v_{il}$ for $j = 1,2..,k$. Suppose we have 5 data, the number of cluster $k = 2$, then we can make the partition matrix. The matrix will be transformed into chromosomes form.
- *Evaluation of fitness value*
  This process evaluates each population by calculating the fitness value of each chromosome or individual. Evaluation of each individual uses a particular function as a performance measure. The Genetic Algorithm intends to maximize the fitness value. The chromosome has a better solution at higher fitness value. The *f* is the fitness value:

$$f = \frac{1}{(h+a)}, \tag{11}$$

h is the objective function described by Equation (10), and a is the small number to avoid division by zero.

- *Selection*
  The selection process is a process to determine which individuals will be selected to perform crossover and mutation processes. In this study, we spin the roulette wheel as much as *N* times [14]. For each repetition, we choose the next population's chromosome. The calculation of probability value P for each individual/chromosome depends on fitness value:

$$P_i = \frac{f_i}{f_{total}} \qquad i = 1,2,\dots,N\ population. \tag{12}$$

  where f_i is the fitness value of i-th individual and f_{total} is the total fitness value of all individuals in a population. Then, we calculate the cumulative probability value P_{cum} and generate a random value of r [0,1] to select the individual. We choose an individual if the cumulative probability $P_{cum} \geq r$.
- *Crossover*
  Crossover is a process for adding diversity or individual variation in a population. We use a one-step Fuzzy K-Prototype to speed up the convergence process, based on the crossover for fuzzy k-modes [15].
  Suppose C1 is a chromosome (fuzzy membership) that has been selected for the crossover process. C1 will be included in the clustering process with only one iteration, so its fuzzy membership will be updated as we explained before in the Fuzzy K-Prototype process on the fifth step, the crossover process.
- *Mutation*
  A mutation is a process of changing the value of one or more genes in a chromosome so that it will produce new individuals. This process aims to avoid local optimum. Gan et al [16] propose that probability for each gene to mutate $P_m \in [0,1]$ is small. If a random number (generated for each gene) is less than P_m $t$he selection of the fuzzy membership is taken place. The mutation process is described following the work of Zhao et all [17].
- *Elitism*
  Elitism is the process of copying an individual or a chromosome that has the highest fitness value. This strategy guarantees that the solution quality obtained by the GA will not decrease from one generation to the next due to the crossover and mutation process.

2.6. Process Flow of Genetic Fuzzy K Prototype Algorithm (GA-FKP Algorithm)

The process flow of Genetic Fuzzy K Prototype is described as follows:

Step 1. Input parameter. The input parameters are data, number of clusters (k), fuzziness coefficient (α), population, maximum generation, and mutation rate. In this study, the number of clusters is determined by the k-prototype algorithm. Then for the value of α is the same as the value of α in a Fuzzy K-Prototype algorithm. The determination of maximum generation depends on the number of iteration when simulating the fuzzy k-prototype algorithm. For example, we do simulation using the FKP algorithm for ten times and the maximum number of iteration is 18, so the maximum generation value can be determined with the value close to 18. For example, we set the maximum generation is 20.

Step 2. Initialization of the population.

Step 3. Evaluation of fitness.

Step 4. Creation of a new population. To generate a new population, we should consider three elements process: selection is the process of selecting the number of chromosomes to include in the next process; crossover in this study using a one-step fuzzy k-prototype as the operator; a mutation is a process to change the value of the gene in a chromosome.

Step 5. Use the new population to run the next process.

Step 6. Stop if the maximum iteration has reached. Back to step three if the maximum iteration has not reached yet.

2.7. Clustering Quality Measure (Evaluation)

To measure the quality of the clustering result, we use several validity indexes depending on the type of data used. The measured dataset without ground truth, we use the Coefficient Variance (CV) index [18]. CV index consists of CU (category utility) for categorical attributes and variance for numeric attributes. In CV, the higher category utility function (CU) is preferable. For numeric attributes, the standard deviation shows values dispersion. Variance is for evaluating the quality of the clustering of numeric data. So, in this case, we expect to get higher CU with the smallest variance. Where CV function is CU divided by variance, it means higher CV is indicating better clustering. We also use the *Partition Coefficient* (*PC*), *Partition Entropy* (*PE*) [19], and the objective function to compare Fuzzy K-Prototype Algorithm and Genetic Algorithm Fuzzy K-Prototype Algorithm because they work for fuzzy clustering only. The higher PC value and the lower both of PE and the objective function value gives a better clustering result. For a dataset with known ground truth, we use purity. The data mining community usually utilizes purity [20], which measures the cluster's "purity" after considering class labels. The higher purity value indicates that we have a better clustering result.

3. Result and Discussion

3.1. Performance of Genetic algorithm-Fuzzy K-prototype

Our experiment is implemented in RStudio software. We use two types of a dataset to test our algorithm: generated data (generated by function in R) and actual data taken from the UCI Repository. As a random initial prototype, K-Prototype(KP), Fuzzy K-Prototype (FKP) is run 20 trials, and we calculate the average accuracy. Also, we set max lte=100, number of intervals T=4, and ε=0.00001 for all experiments. For the Genetic Algorithm Fuzzy K-Prototype (GA-FKP) algorithm, we set the number of population equal to 20, the maximum generation is 50, and the mutation rate is 0.001. For the K-prototype, we use the "kproto" function from the *clustmixtype* package in R. We test our algorithm and compare it to another algorithm in our experiments. In this experiment, we use two datasets: generated data and real-world data.

For generated data, we set up two multivariate data with high correlation and low correlation. This simulation aims to see the robustness of the algorithm proposed in classifying objects with the condition that attributes have strong multicollinearity compared to grouping objects with weak or independent relationships. For real-world data, we use zoo data and credit approval data. Zoo data consists of 101 animals from the zoo. There are 16 attributes with various characteristics to describe the animals, such as animal name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathers, venomous, fins, legs, tail, domestic, and size. We remove animal name attribute for our purpose. The class attributes have seven values: Mammal, Bird, Reptile, Fish, Amphibian, Bug, and Invertebrate. For the credit approval dataset, all attribute names have been converted to meaningless symbols to protect the confidentiality of the data. However, this still suits our purpose since we are not using the dataset to cultivate actual credit screening criteria just for testing our algorithm only. Credit approval data comprises of 690 observations characterized by five numeric attributes, nine categorical attributes, and the class attribute, which has two values.

Table 1. The four indexes of result evaluation for the different clustering algorithm of generated data with a high correlation.

Type of Generated data	Algorithm Type	Coefficient of Variance (CV)	Objective function	Partition Coefficient (PC)	Partition Entropy (PE)
Number of observation (n) = 100, Cluster's number (k) = 7	K-Prototype (KP)	1.0221			
	Fuzzy K-Prototype (FKP)	1.18004	3.11767	0.84555	0.353
	Genetic Algorithm-Fuzzy K-Prototype (GA-FKP)	**1.20423**	**3.1099**	**0.8562**	**0.33764**
Number of observation (n) = 500, Cluster's number (k) = 7	K-Prototype (KP)	0.59281			
	Fuzzy K-Prototype (FKP)	0.7465	14.8379	0.64881	0.6891
	Genetic Algorithm-Fuzzy K-Prototype (GA-FKP)	**0.76724**	**14.7645**	**0.66817**	**0.66158**

Based on Table 1, In data with high correlation, with 100 observations, the Fuzzy K-Prototype algorithm has a CV index of 0.15794 higher than the K-Prototype algorithm. It means that in data with high correlation, Fuzzy can optimize the performance of the K-Prototype algorithm. The determination of centroid for the categorical data proposed in this paper can increase the performance of the clustering method. When we combine the FKP algorithm with the genetic algorithm, the performance of clustering increases again. Based on the result, the GA-FKP algorithm gives better values for all indexes (CV, objective function, PC, and PE). CV value and PC value are higher than the FKP algorithm, while the objective function and PE value of GA-FKP algorithm are lower than the FKP algorithm, with the difference is 0.02419, 0.00777, 0.01065, 0.01536. The PC and PE are used to validate fuzzy membership of fuzzy k-prototype. So, based on this simulation, the genetic algorithm that combines with Fuzzy K Prototype success in optimizing the KP algorithm under the condition that the data has a high correlation or multicollinearity between attributes. When we add the observation to become 500 observations, the performance of the GA-FKP algorithm is also better than the FKP. The difference values between small observation and large observations are 0.02074, -0.0734, 0.01936, and -0.02752. It means GA-FKP is also effective for a larger dataset.

Table 2. The four indexes of result evaluation for the different clustering algorithm of generated data with a low correlation.

Type of Generated data	Algorithm Type	Coefficient of Variance (CV)	Objective function	Partition Coefficient (PC)	Partition Enthropy (PE)
Number of observation (n) = 100, Cluster's number (k) = 7	K-Prototype (KP)	0.33512			
	Fuzzy K-Prototype (FKP)	0.64048	1.53648	0.79689	0.45857
	Genetic Algorithm-Fuzzy K-Prototype (GA-FKP)	**0.64649**	**1.48212**	**0.81126**	**0.42891**
Number of observation (n) = 500, Cluster's number (k) = 7	K-Prototype (KP)	0.30392			
	Fuzzy K-Prototype (FKP)	0.66884	1.26001	0.40372	1.26993
	Genetic Algorithm-Fuzzy K-Prototype (GA-FKP)	**0.66884**	**1.25999**	**0.40376**	**1.26986**

Generally, the result of performing clustering for generated data with a low correlation has the same conclusion with the data with a high correlation. However, an interesting finding in our research appears for the comparative performance of the GA-FKP algorithm between data with high correlation and data with low correlation. The change of the values for four indexes of evaluation shows that the biggest change appears in data with high correlation and larger dataset: 0.02074, -0.0734, 0.01936, and -0.02752. It means that the GA-FKP algorithm is also able to solve the condition of the more complex data with a high correlation between attribute and large dataset.

Table 3. The purity result using different clustering algorithms for zoo data and credit approval data.

Data	Purity		
	K-Prototype (KP)	Fuzzy K-Prototype (FKP)	Genetic Algorithm-Fuzzy K-Prototype (GA-FKP)
Zoo	0,84851	0,93762	**0,94059**
Credit Approval	0,71732	0,85617	**0,86232**

To test our proposed algorithm, we compare the purity index result of a comparative study on real-world data. We expect to get a higher value of purity. In Zoo data, the purity of FKP is 0.08911 higher than the KP algorithm, and then when we apply GA to FKP algorithm, the purity increase 0.00297. In credit approval data, the purity of GAFKP is 0.00615 higher than FKP and 0.145 higher than KP Algorithm. It means that GA-FKP also gives a better result than the KP and the FKP algorithm no only in generated data conditioned with a high and low correlation, but also in the real-world data.

3.2. The novelty of research

The novelty of our research that different with prior researchs that only simulation the algorithm with real world dataset, we generate some data simulation in different condition of correlation rate between attributes in a dataset and different size of observation. Based on table 1and 2, Genetic Algorithm- Fuzzy K-Prototype algorithm (GA-FKP algorithm) not only solves the weaknesses of K-prototype algorithm (KP algorithm) about the local optimum solution and the weaknesses of the use mode as center for categorical attributes, but the simulation of generated data that we implication to whole algorithm show that our propose algorithm, GA-FKP algorithm, is still even more effective for dataset with high correlation between attributes and has larger datasets. Therefore cost time proceesing with this algorithm is lower than prior research.

4. Conclusion

The Genetic Algorithm-Fuzzy K-Prototype can resolve the two weaknesses of the K-Prototype Algorithm. The Fuzzy centroid can solve the weaknesses of the cluster center for categorical attributes, and the Genetica Algorithm can optimize the solution of the K-Prototype Algorithm. The algorithm is robust for the data with a high dimension. The Genetica Algorithm hybrid with Fuzzy K-Prototype Algorithm optimizes the process of clustering for the data that has a high correlation between attribute and also has a large dataset. The result is again based on the generated data to train our proposed algorithm high and low correlation, Finally, when we implemented the algorithm to the real-world dataset, the performance of this algorithm is unchanged, still powerful than the K-Prototype algorithm.

Acknowledgements

We thank the Center for Research and Community Service (PPPM) , STIS Polytechnic of Statistics, Jakarta, for the support in publishing this work.

References

1. Sumanthi, S., & Sivanandam, S. Introduction to Data Mining Principles, Studies in Computational Intelligence (SCI). In.; 2016. p. 1-23.
2. Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. San Francisco, CA, itd: Morgan Kaufmann.; 2012.
3. Wang L,JH,&GX. Image Segmentation by a Robust Clustering Algorithm. In.: Communication; 2004. p. 74-81.
4. Inza Ia,CB,ARA,BE,APL,&LJA. Machine Learning: An Indispensable Tool in Bioinformatics. In.: Bioinformatics Methods in Clinical Research; 2010. p. 593.
5. Flasi'nski M. Introduction to Artificial Intelligence. Pattern Recognition and Cluster Analysis. 2016;: p. 141–156.
6. Ahmad A,&DL. Data and Knowledge Engineering. A k-mean clustering algorithm for mixed numeric and categorical data. 2007; 63(2): p. 503-527.
7. Li C,&BG. IEEE Transactions on Knowledge and Data Engineering. Unsupervised learning with mixed numeric and nominal data. 2002; 14(4): p. 673-690.
8. Lam D,WM,&WD. Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning. IEEE Access. 2015; 3(c): p. 1605-1616.
9. Huang Z. Clustering Large Data Sets with Mixed Numeric and Categorical Values; 1997.
10. Kim DW,LKH,&LD. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters. 2004; 25(11): p. 1263-1271.
11. Ji J,PW,ZC,HX,&WZ. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowledge-Based Systems. ; 30: p. 129–135.
12. Nooraeni R. Cluster Method Using a Combination of Cluster Kprototype Algorithm and Genetic Algorithm for Mixed Data. Jurnal Aplikasi Statistika & Komputasi Statistik. 2016;: p. 81–97.
13. Ze Dong HJML. An Adaptive Multiobjective Genetic Algorithm with Fuzzy c_means for Automatic Data Clustering. Hindawi. 2018 May; 2018
14. Basak J,DRK,&PSK. Unsupervised feature selection using a neuro-fuzzy approach. Pattern Recognition Letters. 1998 June; 19: p. 997-1006.
15. Zadeh LA. A fuzzy-set-theoretic interpretation of linguistic hedges. Journal of Cybernetics. 1972; 2(3): p. 4-34.
16. Gan G,WJ,&YZ. A genetic fuzzy k-Modes algorithm for clustering categorical data. Expert Systems with Applications. 2009; 36: p. 1615–1620.
17. Zhao L,TY,&GM. Genetic Algorithm. In.; 1996. p. 716–719.
18. Hsu CC,&HYP. Incremental clustering of mixed data based on distance hierarchy. Expert Systems with Applications. 2008; 35(3): p. 1177–1185.
19. Dagher I. Complex fuzzy c-means algorithm. Artificial Intelligence Review. 2012; 38(1): p. 25–39.
20. Zhao Y,&KG. Criterion functions for document clustering: Experiments and analysis. Machine Learning. 2004; 55(3): p. 311–331.