Avrupa Bilim ve Teknoloji Dergisi
Cilt. 5, No. 9, S. 28-37, Aralık 2016
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Vol. 5, No. 9, pp. 28-37, December 2016
Copyright © EJOSAT
Research Article

A New Clustering Algorithm of Hybrid Data According to Weights of Attributes

Osman COREKCI[1], Ayla SAYLI[2*]

[1] Yildiz Technical University, Department of Mathematical Engineering, Istanbul 34010, Turkey; osmancorekci.math@gmail.com
[2] Yildiz Technical University, Department of Mathematical Engineering, Istanbul 34010, Turkey; sayli@yildiz.edu.tr

(First received 10 October 2016 and in final form 22 November 2016)

Abstract

Separating large data into similar clusters is one of the basic problems of data mining. Storing large data in an organized way has currently increased the importance of the methods developed for clustering. Even if the hierarchical clustering methods give effective results, they are still inadequate due to their computational complexity. Non-hierarchical clustering methods cannot be used for all data types because of the cost function which cannot run by categorical data. Recently, some non-hierarchical clustering methods have been improved for categorical and hybrid data. In addition, the weights of attributes in clustering might be different due to the nature of the data or the expected results. In this paper, we introduce an algorithm which has been improved for the clustering of large hybrid data in an effective way that also includes the weights of attributes. This algorithm, mainly based on the K-Prototypes algorithm, will be called "W-K-Prototypes". The computational results show that the algorithm can be used efficiently for clustering.

Keywords: Data mining; clustering; clustering analysis; clustering methods; clustering hybrid data; K-Prototypes algorithm; weighted attributes

1. Introduction

Extracting meaningful information from large data is currently one of the most popular research fields. Large data can be stored by means of continuously improving technology and can be used to evaluate the past and predict the future, which is regarded as Data Mining. One of the most important methods of this process of data mining is clustering analysis. Clustering is a process of separating objects in such a way that objects with similar characteristics should be in the same group. Clustering analysis is a process of dividing the data into groups whose labels were not defined a priori. After this process, the obtained clusters show high homogeneity for intra-cluster and high heterogeneity for inter-cluster [7].

Clustering analysis was first used by Tyron in 1939 [9]. After the 1980s, its usage became widespread. The book written by Robert Sokal and Peter Sneath named "The Basics of Numeric Classification Knowledge" has been an important milestone in the field [8]. As a result of these studies, today we have different clustering analysis methods.

However, as time goes on, data becomes larger and data types vary. That is the reason why current clustering analysis methods are inadequate. The hierarchical clustering method is the most well-known clustering analysis method which is able to cluster data which has both numerical and categorical characteristics, but it loses its usefulness due to high computational complexity [1]. Non-hierarchical clustering methods based on K-Means have reasonable calculation complexity. These methods need a cost function which requires a Euclidean measure and cluster average, and they cannot run with categorical data [8]. At this point, in recent studies, the researchers were encouraged to develop a new non-hierarchical method based on K-Means which can be used with the categorical data. A method named K-Modes was proposed by Huang in order to cluster solely categorical data with a non-hierarchical method [2]. This method changes the average value at the step of defining the cluster center with the mode value and uses δ (overlap) instead of Euclidean distance measure. In the algorithm named K-Prototypes, Huang offered an effective solution for the clustering problem of hybrid data based on the logic of the K-Means algorithm in 1997 [1,3]. We will also analyze Huang's study in this article.

The next section introduces the mathematical preliminaries needed for this study. In the third section, the K-Prototypes algorithm by Zhexue Huang will be given for the hybrid data. In the fourth section, the version of W-K-Prototypes will be described in order to be able to account for the weights of attributes. In the fifth section, these two algorithms will be tested on the data and computational results will be shown. In the last section, we will explain the conclusions and future work.

2. Mathematical Preliminaries

The data consisting of n objects is represented by $X = \{ x_i \mid 0 < i \leq n \}$ where every $x_i = \{ x_{ij} \mid 0 < j \leq m \}$ is a vector composed of m attributes and can be openly shown as $x_i = [x_{i1}, x_{i2}, \dots, x_{im}] \in X$ notation. In non-hierarchical clustering

[2] Correspondence: sayli@yildiz.edu.tr; Tel.: +90-533-475-0667

analysis, the problem is to separate this X into clusters according to the number of k no-predetermined groups. It is required to have an objective criterion. That criterion enables us to define the quality of the clustering process. For that, the function used is called the cost function [4]. This function is formulated as

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(x_i, Q_l) \qquad (1)$$

Here, n represents the number of elements for cluster X, where $x_i \in X$ and Q_l is the center of the cluster l. $y_{i,l}$ is an element of the matrix partitioning. The elements of the partitioning matrix, Y_{nxl} are respectively 1 or 0 according to whether or not the l element belongs to the n cluster, and d is the distance measure of Squared Euclidean, unless otherwise stated. $y_{i,l}$ is the value in the interval 0 and 1 and $\sum_{l=1}^{k} y_{i,l} = 1$. If $y_{i,l} \in \{0,1\}$, then Y_{nxl} is named as the *hard partition*; if $y_{i,l} \in [0,1]$, then Y_{nxl} is named as the f*uzzy partition*. In this document, we will deal with the hard partitioning. Hence, an object can only belong to one cluster.

The cluster center, $Q_l = \{q_{l1}, q_{l2}, .., q_{lm}\}$ is calculated by the following:

$$q_{lj} = \frac{1}{n_l} \sum_{i=1}^{n} y_{il} x_{ij} \qquad (2)$$

Here n_l is the number of elements of the l cluster. In non-hierarchical clustering methods, we need to have the distance metric to determine the distance between each cluster and the cluster centers to represent the cluster objects. The measure of the Squared Euclidean dimension is used only for the cluster vectors including numerical characteristics. The average of the vectors in the cluster is taken as the center as representative. The measure proposed by Huang is the following function for the categorical data.

$$d(x_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \qquad (3)$$

Here, $\delta(x, y)$ is a function that has a value of 1 if its categorical variables are the same; if not, the value becomes 0. x_{ij}^r and q_{lj}^r are numerical, and x_{ij}^c and q_{lj}^c are categorical. m_r is the number of numerical attributes and m_c is the number of categorical attributes. γ_l is the weight of the categorical features.

This distance measure is used when the cost function can be rewritten in the following way [1]:

$$\begin{aligned} E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 \\ + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \\ = E_l^r + E_l^c \end{aligned} \qquad (4)$$

Here E_l^r is the cost function of the numerical attributes and E_l^c is the cost function of the categorical attributes. Given this cost function, then in order to minimize E_l^c, the cost function and categorical features of the center need to be chosen from the most frequent values because in the case where the sigma function is the same, it gives a value of 0 and decreases the contribution to the sum. That is why Huang proposed the following method: the average of the numerical variables and the frequency of the categorical variables should be chosen.

The γ_l coefficient in the equation of (3) has been proposed with the aim of defining the weight at the measure of the dimension of the categorical data. For example, when the numerical value ranges of the variables are very wide and if the categorical variables are different, these add a value of 1 to the distance. This value may remain meaningless and small. On the contrary, the numerical values in the range [0.5, 0.6], with changes across a small range, play a dominant role in determining the value of 1 if the offset creates an undesirable situation. In order to preclude this case, Huang proposed using the value of the standard deviation σ_l (the average of the numerical properties of the standard deviations in the set) for the purpose of determining the value of γ_l. Therefore, in the case where the categorical variables are different, there will hardly be any effect on the measure of the dimension as in the case where the numerical values are different. This proposal should not be expected to always give better results but when all the data are taken into consideration, this can be accepted as a good solution. Huang shared some results for the different values of γ_l.

2.1. Standardization of data

There are various methods about standardization of the variables. One of these methods is the most well-known referred to as Z-Score standardization. With this method, the data average turns to 0 and the standard deviation turns to 1. Thus, whether the data attributes are above or below the average can be determined just by looking at its value. New values of the attributes in the Z-Score standardization are calculated with the following formula.

$$x' = \frac{x - mean(X)}{sd(X)} \qquad (5)$$

Here $mean(X)$ is the average of the data cluster X, and $sd(X)$ is the standard deviation of the data cluster X. Even if the Z-Score method turns the data into unknowns, it still gives the best result in distance calculation. After the clustering process, in order to find out which data belongs to what cluster, the data should be applied to a process with an ID value; not putting this ID value in the distance calculation is also a method.

In hybrid data, it is enough to standardize the numerical part of the data. It is obvious that the categorical data cannot be standardized through the Z-Score method because of the different distance measures used; if they are not standardized, there is no problem.

2.2. K-Prototypes algorithm

K-Prototypes algorithm has the following similar steps as K-Means.

[i] Choose k as the initial center (Prototype) in order to represent the clusters in the data set X.

[ii] Assign each object in the data set X to the cluster including the nearest prototype.

[iii] Re-compute the prototypes of the clusters after the assignment of all objects, then assign each object in the data set X to the cluster including the nearest prototype.

[iv] Re-iterate the step (iii) till no object is unplaced.

The distance measure used in this algorithm is given in the equation (3). Recalculating the weight γ_l for every step and every cluster is one disadvantage of this algorithm.

In fact, the distance measure used in the algorithm consists of the sum of the results of two number distance measures which are the Squared Euclidean Distance measures for the numerical values and overlap for the categorical values. Alternative measures may be considered in order to make the distance measure used in K-Prototypes more modular.

3. W-K-Prototypes

Sometimes the attributes of the data might not have equal importance. The significance level may differ by the scope of the requested information extracted from the data. For example, when the data held is about student information then if a classification is going to be made based on the achievement status of students, it is expected that the attributes such as weight, length, and school number have less impact. Those attributes may even be removed from the data. In the case of the classification being based on the physical similarities of students, it is expected that marks have much less impact on the clustering process. Besides that, regardless of the clustering process, in the nature of some data, weight distribution might be encountered. Such situations should be taken into consideration for clustering [6].

The weight γ_l, distance measure used in Huang's K-Prototypes Algorithm, is the weight in the context of the influence of all categorical data, but not a weight mentioned above. It is a serious burden for the algorithm to recalculate γ_l again in every step and for every cluster. Huang, in his study, uses only one γ value during the processing of the algorithm test instead of calculating a value of γ_l for every step. In the algorithm named W-K Prototypes, there is no need to have such a γ density as the distance measure amongst objects. Before the data is clustered with a W-K Prototypes algorithm, we need to standardize the data with the Z-Score method thereby removing both the need of this coefficient and the problem of calculating the values of γ. Due to Z-Score normalization, since the value of the average standard deviation will be 1, indeed it can be considered as if the γ coefficient is 1 for every step.

$$d(x,y)=\sum_{i=1}^{m_{num}} w_i(x_i^r-y_i^r)^2+\sum_{i=1}^{m_{cat}} w_i\delta(x_i^c,y_i^c) \qquad (6)$$

-where w_i is the weight of the i'th attribute, and for both categorical and numerical types of all attributes $\sum_{i=1}^{m} w_i = 1$. The distance function used for distances of the categorical data is defined as follows:

$$\delta(x,y)=\begin{cases}0 & if\ x=y\\ 1 & else\end{cases} \qquad (7)$$

The Algorithm steps for the distance measure between the two objects are the same for the K-Prototypes Algorithm.

(i) Choose k, the initial center (Prototype) in order to represent the clusters in the data set X.

(ii) Assign each object in the data set X to the cluster including the nearest prototype.

(iii) Re-compute the prototypes of the clusters after the assignment of all objects, then assign each object in the data set X to the cluster including the nearest prototype.

(iv) Re-iterate the step (iii) till no object is unplaced.

4. Computational Results

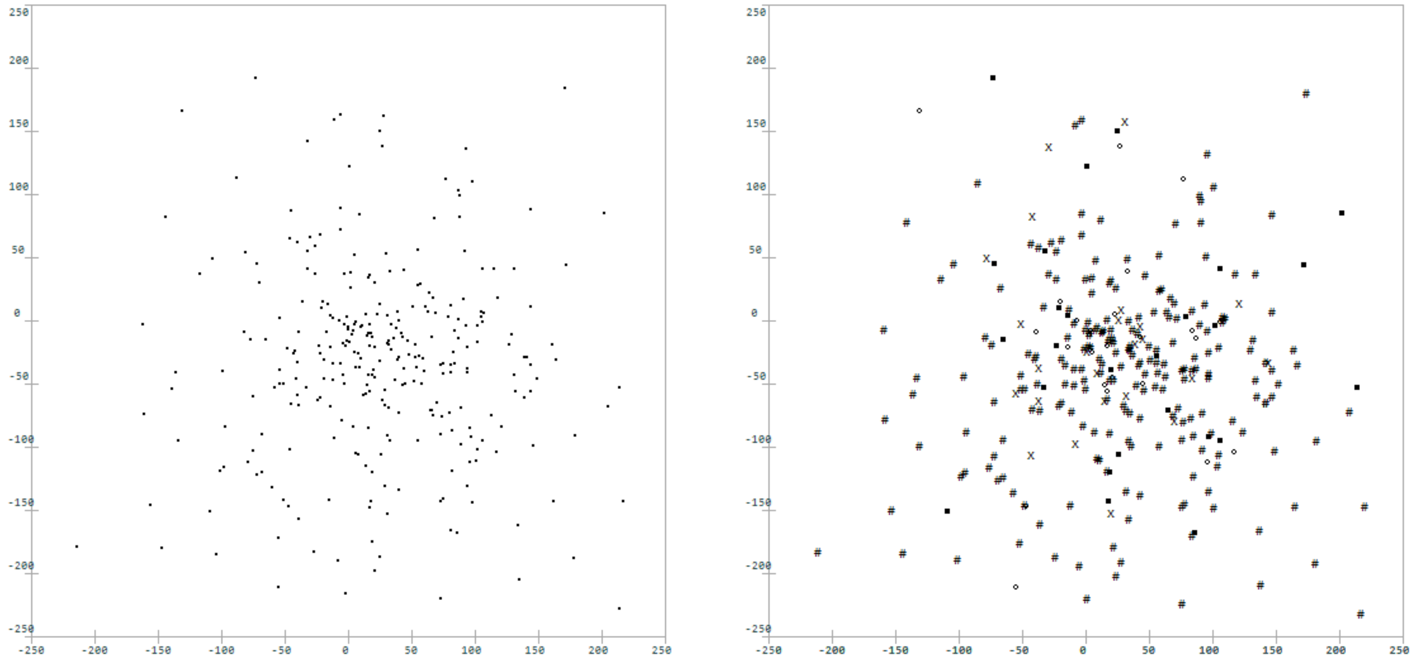
We artificially generate the data for comparison of the W-K-Prototypes Algorithm and K-Prototypes Algorithm in order to replicate generation of the artificial data that Huang used in his study [1]. The distributions of the generated data are shown in Figure 1.a. and Figure 2.a. After adding the categorical third dimension data to the data set, the new distribution can be seen in Figure 1.b. and Figure 2.b. Clustering results raised by clustering for the different weight values are shown in Figure 3. Moreover, the cluster number, the k value, object number and the effect of the changes of these inputs are also shown in Table 1 and Figure 5.

The computer used for the tests of these algorithms had Intel Core 7 processor, 32 GB memory, Windows 8.1 operating system and 8 cores.

Inputs of our algorithm are the k number of clusters, X data set and w=[w_1,w_2,...,w_m], and the weight vector is such that each element corresponds to each attribute. Results for various k-value and w-weight vectors of the two synthetic data sets created are given in Figure 3 and Figure 4. As a first example, let us deal with the results for the X-data set in Figure 3. The numerical attributes of this data set are configured in the union of the sets that have 4 normal distributions in two dimensions. Therefore, let us acquire the results for k=4 and different weight values.

When the graphs below, which are used in the representation of the results, are examined, for weight vector w=[0.33,0.33,0.34] (Figure 3.a.), the weights are set to equal value and the sets are split into four with normal distribution. For the weight vector w=[0.2,0.2,0.6] (Figure 3.e.), the first dimension outweighs the others. As expected, the set is observed to have a clustering on the horizontal axis. As for weight vector w=[0.8,0.1,0.1], the 1st dimension is taken into focus and, expectedly, the output of the clustering appears in a scattering throughout the horizontal axis. For weight vector w=[0.1,0.8,0.1] (Figure 3.c.), the 2nd dimension dominates the others, and similar to the previous result, the set is observed to have a clustering on the vertical axis. While setting the weight vector, deleting the 1st component has no effect. In other words, the result of assigning 0 to the 1st component w=[0.0,0.5,0.5], is shown in Figure 3.f. The distribution, the result of setting w=[0.1,0.1,0.8], shows that one trait outweighs the others, as shown in Figure 3.d.

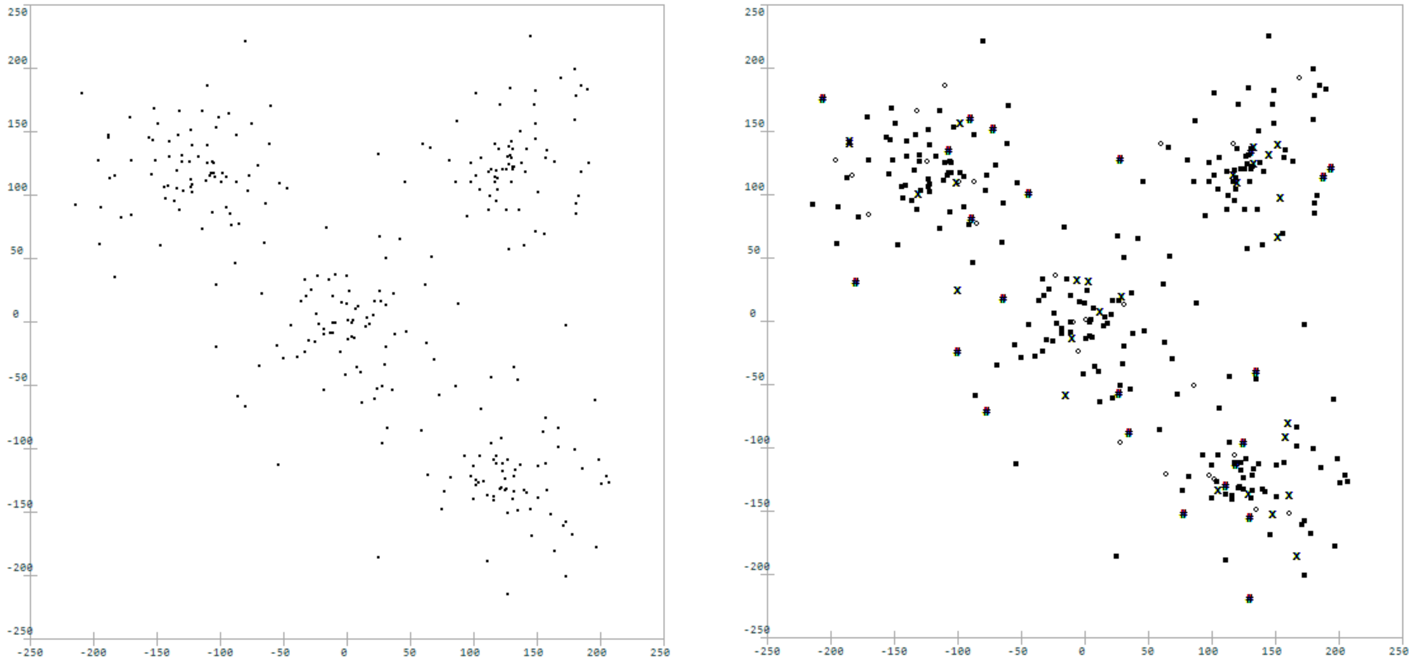
Similar to the results given in Figure 2 and for various weight vectors, as expected, the results of clustering are observed depending upon a dominant attribute. In Figure 4.a. the results of clustering for w=[0.33,0.33,0.34] are given and uniformly clustered sets are observed. In Graph 4.e., the results of clustering for w=[0.2,0.2,0.6] are as expected: categorical data prevails to define the cluster of two objects that are closely related to each other with respect to the first two attributes. As for Figure 4.b. and Figure 4.c., as in the data of our first experiment, the 1st and 2nd attributes behave in a dominant manner and it is observed that clusters are formed by objects grouped, first horizontally, and then vertically.



(a)

(b)

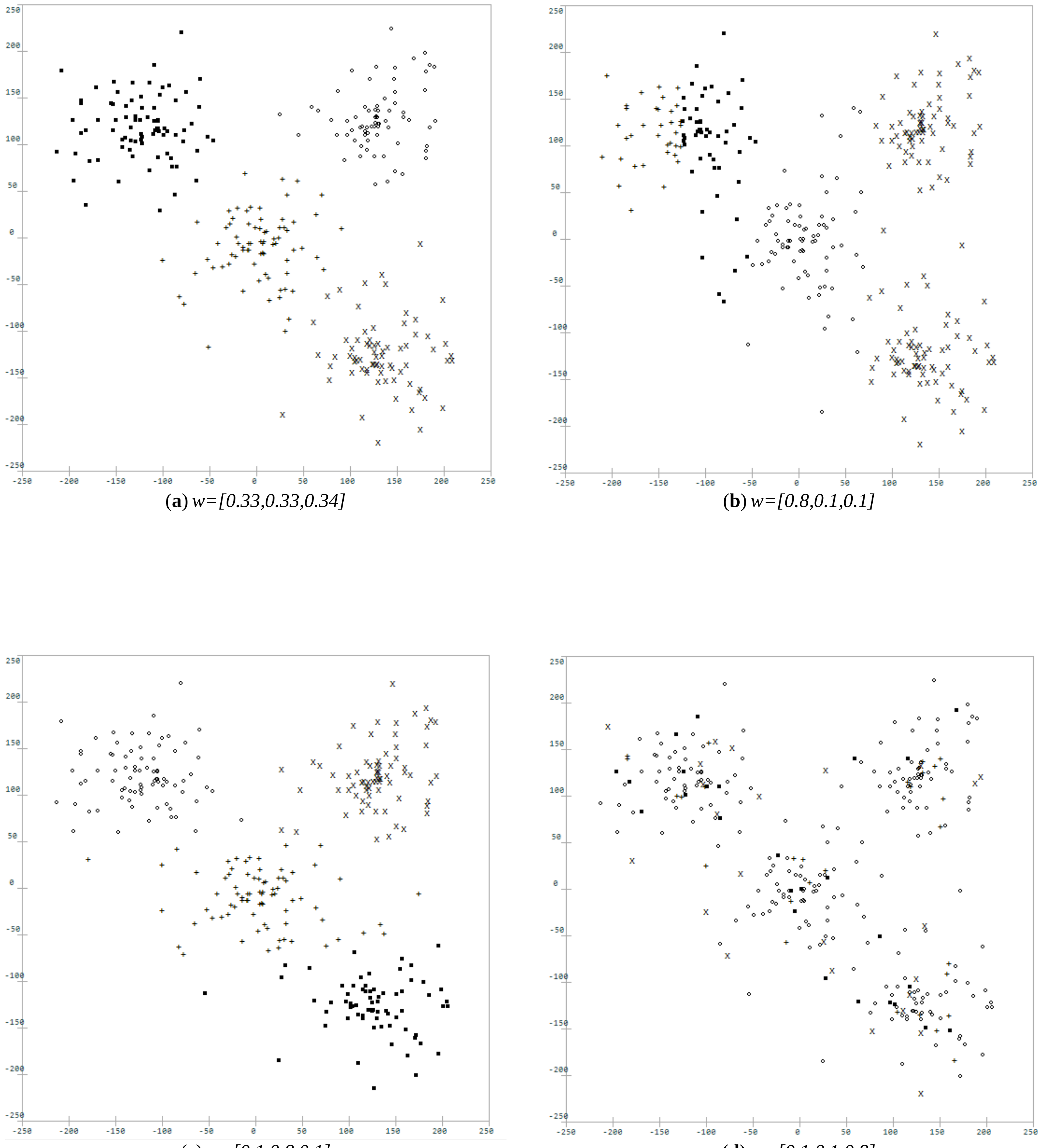
Figure 1. *Sample with one normal distribution.*



(a)

(b)

Figure 2. *Sample with four normal distributions.*

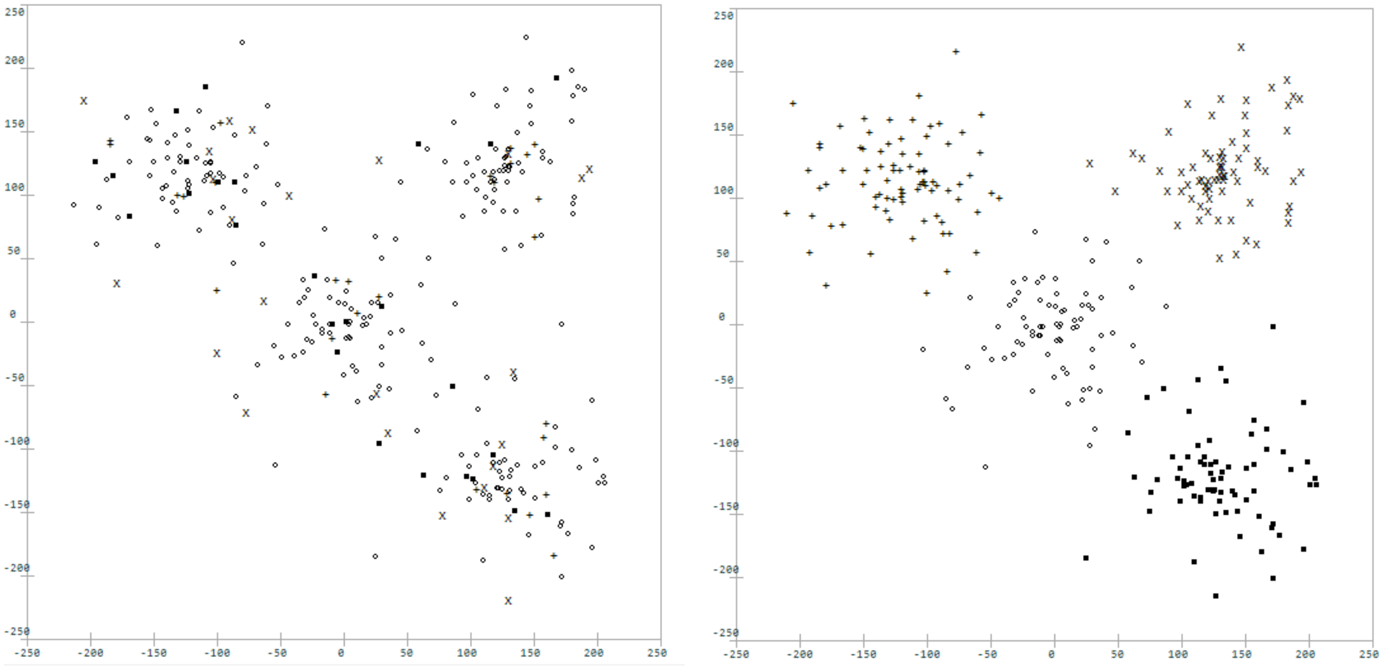


(**a**) *w=[0.33,0.33,0.34]*

(**b**) *w=[0.8,0.1,0.1]*

(**c**) *w=[0.1,0.8,0.1]*

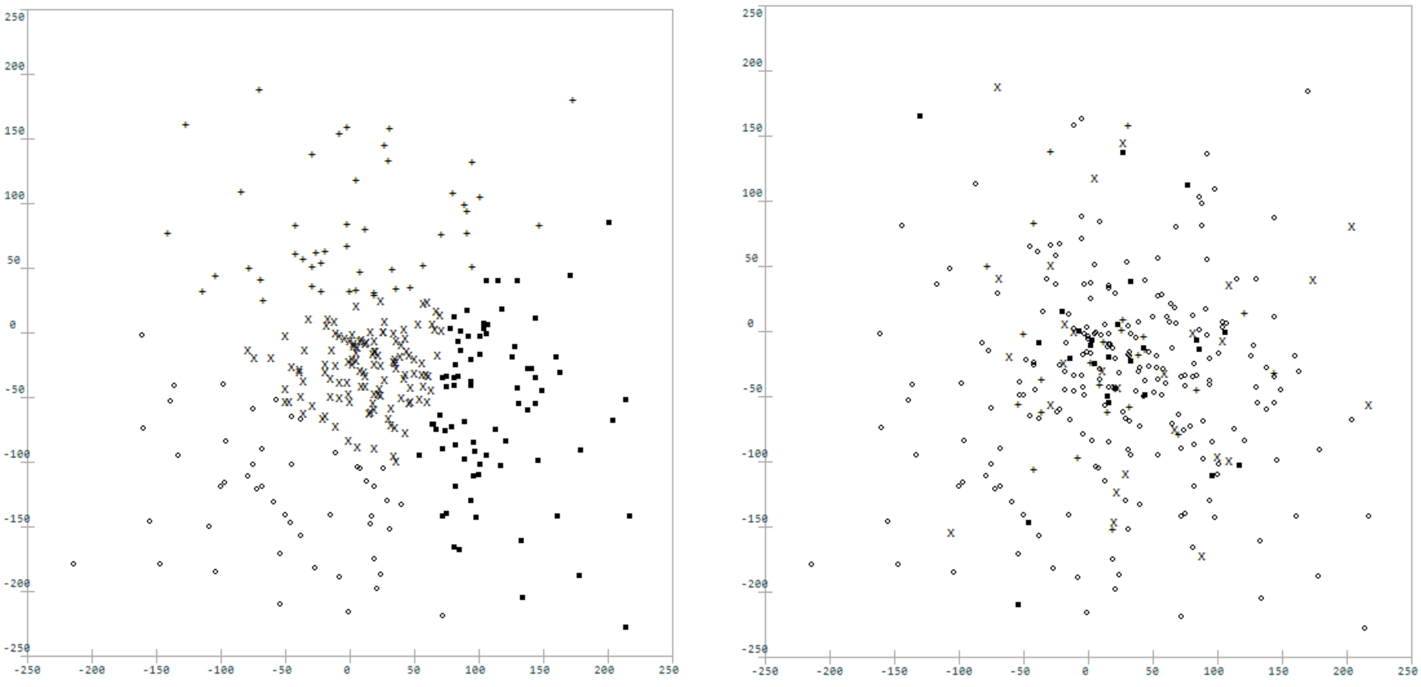
(**d**) *w=[0.1,0.1,0.8]*



(**e**) *w=[0.2,0.2,0.6]*

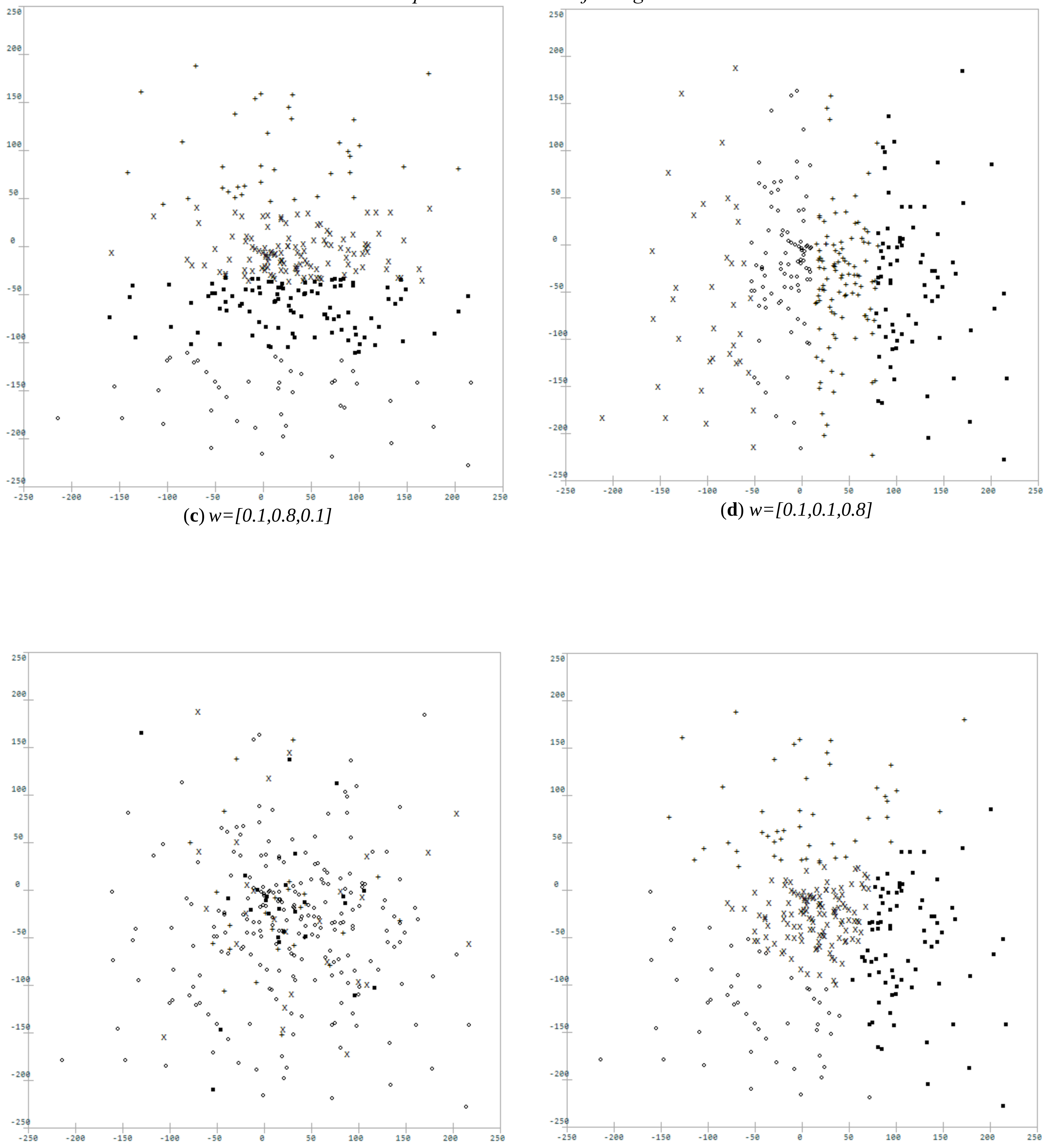
(**f**) *w=[0.0,0.5,0.5]*

Figure 3. *Clustering results for different weight vectors.*



(**a**) *w=[0.33,0.33,0.34]*

(**b**) *w=[0.8,0.1,0.1]*



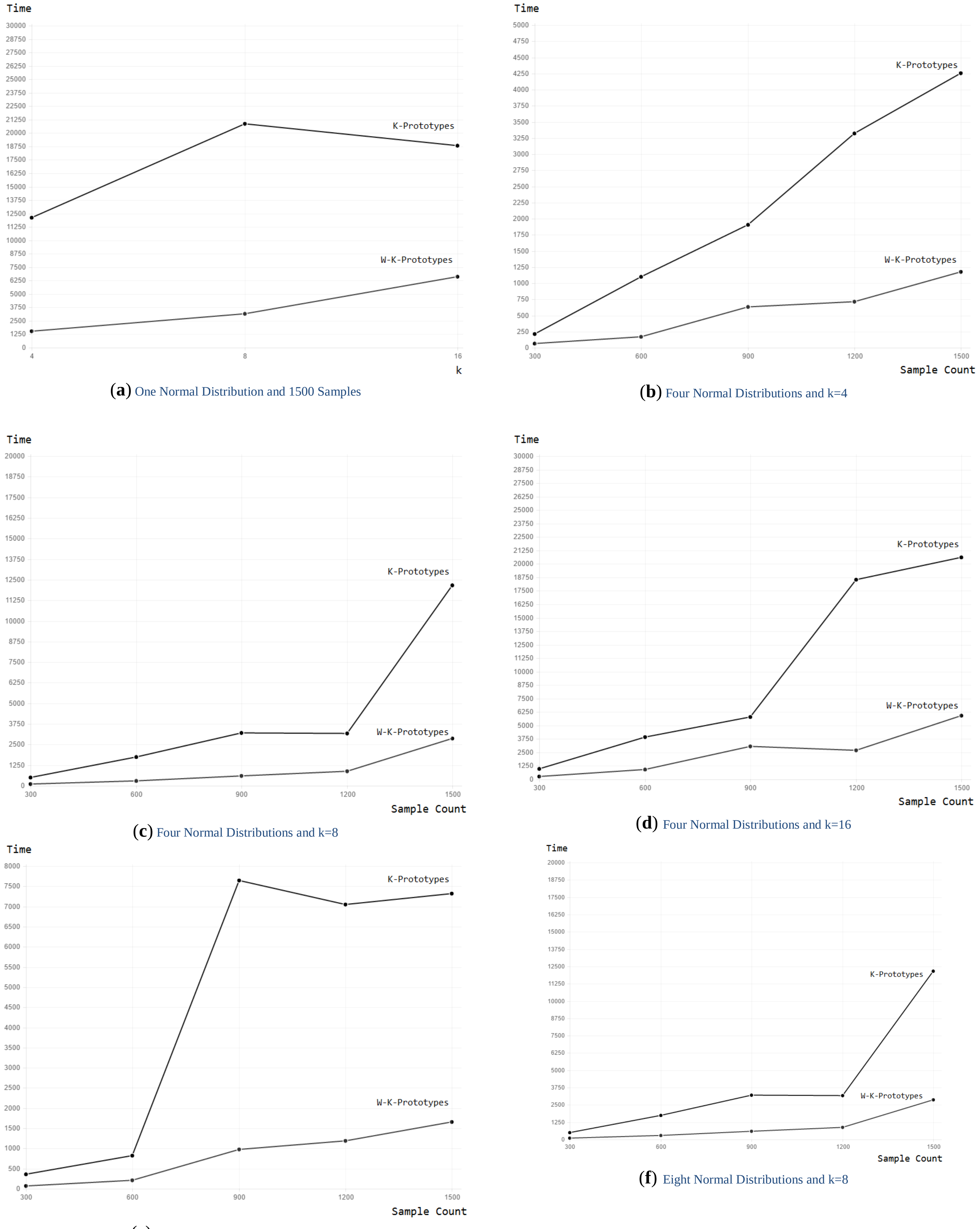
(**c**) *w=[0.1,0.8,0.1]*

(**d**) *w=[0.1,0.1,0.8]*

(**e**) *w=[0.2,0.2,0.6]*

(**f**) *w=[0.0,0.5,0.5]*

Figure 4. *Clustering results for different weight vectors.*



(a) One Normal Distribution and 1500 Samples

(b) Four Normal Distributions and k=4

(c) Four Normal Distributions and k=8

(d) Four Normal Distributions and k=16

(e) Eight Normal Distributions and k=4

(f) Eight Normal Distributions and k=8

Figure 5. *Comparisons between the Algorithms of K-Prototypes and W-K-Prototypes.*

Table 1. *Test results with different inputs for K-Prototypes and W-K-Prototypes.*

Cluster Count	Sample Count	k	Iteration Count		Estimation Time (milliseconds)	
			K-Prototypes	W-K-Prototypes	K-Prototypes	W-K-Prototypes
4	300	4	8	12	212	64
4	600	4	12	7	1099	169
4	900	4	10	14	1905	632
4	1200	4	10	9	3321	713
4	1500	4	8	10	4254	1175
8	300	8	18	14	499	100
8	600	8	18	12	1742	293
8	900	8	17	14	3203	594
8	1200	8	9	10	3171	876
8	1500	8	23	20	12158	2866
4	300	8	13	13	342	84
4	600	8	19	22	1908	550
4	900	8	19	8	3854	443
4	1200	8	15	24	5418	2459
4	1500	8	34	20	19322	2646
8	300	4	15	11	359	68
8	600	4	10	9	821	209
8	900	4	40	23	7644	974
8	1200	4	22	15	7048	1187
8	1500	4	14	12	7319	1656
4	300	16	32	34	975	263
4	600	16	37	34	3921	915
4	900	16	27	56	5790	3059
4	1200	16	51	33	18534	2692
4	1500	16	36	42	20608	5915
1	1500	4	22	13	12108	1527
1	1500	8	38	21	20851	3154
1	1500	16	32	44	18817	6609
1	3000	16	63	26	138977	15029

Taking into account the results in *Table 1* and *Figure 5*, it can be seen that the W-K-Prototypes Algorithm is obviously superior to the K-Prototypes Algorithm in terms of the processing time. It is observed that the W-K-Prototypes Algorithm runs five times faster with respect to the means of the rates of the processing times. As detected in *Figure 5.e.*, it can be seen that the W-K-Prototypes Algorithm gives more efficient results particularly in the sense of clustering the sets, with normal distribution, into relatively fewer numbers of sets.

When the results of the clustering are analyzed, it is observed that the K-Prototypes and W-K-Prototypes algorithms produce almost the same clusters. The processing time of the W-K-Prototypes Algorithm in *Table 1* includes the time of the Z-score and inverse Z-Score processing.

4. Conclusions and Future Works

In this paper, we introduce an algorithm which integrates the various attributes of the data into the clustering analyzes by means of various weights in the hybrid data with huge dimensions. This algorithm is named the W-K-Prototypes Algorithm because it is fundamentally based on the K-Prototypes algorithm and efficaciously runs a clustering process by means of a non-hierarchical method without the necessity of the coefficient γ_l. In the making of the algorithm, the distance measures used in the K-Prototypes Algorithm; the method for determining the center of the cluster; and the coefficient γ_l on the effect of the categorical data on the distance, are considered independently from each other and this is the starting point of this approach. Similarly, the components used, depending upon the scope of the expected value and of the data, can be applied in future studies. For instance, in order to compute the distance of the categorical data, a measure based on the value of the frequency of the relevant attribute can be used or, instead of the Euclidean distance measure, another distance measure can be applied in order to increase the degree of the clustering quality [5]. It is obvious that Z-score standardization, the process used for application of the W-K-Prototypes Algorithm, possesses a disadvantage for huge data. Instead, studies on efficiently determining the coefficient γ can be performed.

References

Huang, Z. (1997a) Clustering Large Data Sets with Mixed Numeric and Categorical Values, In Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific.

Huang, Z. (1997b) A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, The University of British Columbia, Canada, pp. 1–8.

Huang, Z. (1998) “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”, Data Mining and Knowledge Discovery, Vol. 2, No. 3, pp. 283 – 304.

Anderberg, M. R. (1973) Cluster Analysis for Applications, Academic Press.

S. Boriah, V. Chandola and V. Kumar. (2008) “Similarity Measures for Categorical Data: A Comparative Evaluation”, Proceedings of the 8th SIAM International Conference on Data Mining, pp. 243 – 254.

Renato Cordeiro de Amorim. (2015) A survey on feature weighting based K-Means algorithms.

Kantardzic, M. (2003) Data Mining: Concepts, Models and Algorithms, IEEE Press and John Wiley, New York.

B. Everitt. (1974) Cluster Analysis. Heinemann Educational Books Ltd.

Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.