Received November 21, 2020, accepted February 1, 2021, date of publication February 4, 2021, date of current version February 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057113

A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data

JINCHAO JI[1,2,3], RUONAN LI[1,2,3], WEI PANG[4,5], FEI HE[1,2,3], GUOZHONG FENG[1,2,3], AND XIAOWEI ZHAO[1,2,3]

[1]School of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2]Institute of Computational Biology, Northeast Normal University, Changchun 130117, China
[3]Key Lab of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China
[4]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, U.K.
[5]Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing, Xi'an University of Technology, Xi'an 710048, China

Corresponding authors: Wei Pang (w.pang@hw.ac.uk) and Xiaowei Zhao (zhaoxw303@nenu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61502093, Grant 61802057, Grant 61977015, and Grant 62077012; in part by the Fundamental Research Funds for the Central Universities under Grant 2412019FZ048, Grant 2412019FZ047, Grant 2412019FZ052, Grant 2412019ZD014, and Grant 2412020XK003; and in part by the Science and Technology Research Project of "13th Five-Year" of the Education Department, Jilin, under Grant JJKH20190290KJ.

ABSTRACT Clustering data with both numeric and categorical attributes is of great importance as such data are ubiquitous in real-world problems. Multi-view learning approaches have proven to be more effective and having better generalisation ability compared to single-view learning in many problems. However, most of the existing clustering algorithms developed for mixed numeric and categorical data are single-view. In this research, we propose a novel multi-view clustering algorithm based on the k-prototypes (which we term Multi-view K-Prototypes) for clustering mixed data. To the best of our knowledge, our proposed Multi-view K-Prototypes is the first multi-view version of the well-known k-prototypes algorithm. To cluster the mixed data over multiple views, we present a novel representation prototype of cluster centres in the scenario of multiple views, and we also devise formulas for updating the cluster centres over each view. Then we propose the concept of consensus cluster centres to output the final clustering result. Finally, we carried out a series of experiments on four benchmark datasets to assess the performance of the proposed Multi-view K-Prototypes clustering. Experimental results show that the Multi-view K-Prototypes algorithm outperforms the seven state-of-the-art algorithms in most cases.

INDEX TERMS Data clustering, multi-view learning, mixed data, numeric and categorical attributes.

I. INTRODUCTION

Clustering analysis, which identifies the nature groups of data objects in an unsupervised manner, is a fundamental task in data mining and machine learning [1]–[4]. Clustering algorithms have been widely used in information retrieval [5], [6], privacy preserving [7], social media analysis [8], text analysis [9], image analysis [10], bioinformatics [11], [12], and sentiment analysis [13] etc. Clustering analysis aims to divide the data objects with similar characteristics into the same clusters, and the ones with dissimilar characteristics into different clusters [14]. The existing clustering algorithms in the literature can be classified into two types: partitional and hierarchical [2]. The partitional clustering algorithms allocate data objects within a dataset into a predefined number of clusters by optimizing an objective cost function, whereas the hierarchical clustering algorithms divide these data objects into a dendrogram of partitions by utilizing an agglomerative or divisive strategy [15].

In partitional clustering algorithms, the k-means algorithm is widely used in many fields due to its simple and efficiency [16]. To deal with the fuzziness of data objects in clusters, the fuzzy k-means algorithm was proposed by Bezdek, Ehrlich, and Full [17]. These two algorithms were developed for the numeric datasets. In many applications, data objects with both numeric and categorical attributes are commonly encountered. These two types of attributes have different domains of values, and they coexist in the clusters and data objects. Thus, how to represent and update the centre of a cluster and design appropriate dissimilarity measures between the centre of a cluster and a data object are the main challenges in clustering mixed data. Several algorithms have

The associate editor coordinating the review of this manuscript and approving it for publication was Tossapon Boongoen.

been proposed for clustering the mixed data. Huang [18] proposed the well-known k-prototypes algorithm, which combined the k-means approach with the k-modes approach, to cluster mixed data. Bezdek *et al* presented the fuzzy k-prototypes algorithm to deal with the uncertainty about which cluster a data object belongs to [19]. The k-prototypes algorithm and its fuzzy version are extended in [14], [20], [21] by utilizing the influence of attributes and improving the cluster centre representation. Li and Biswas developed a hierarchical approach SBAC (Similarity-Based Agglomerative Clustering) by using Goodall similarity [22]. Hsu and Chen presented the method CAVE (Clustering Algorithm based on the Variance and Entropy), which requires to construct the distance hierarchy for categorical attributes [23]. Lam, Wei, and Wunsch introduced the UFLA method by integrating the fuzzy ART (adaptive resonance theory) with the UFL (unsupervised feature learning) [24]. Zheng *et al.* developed the approach EKP (Evolutionary k-prototypes) by using the framework of evolutionary algorithm [25]. David and Averbuch introduced the approach SpectralCAT, which transforms the numeric attributes into the categorical ones [26]. Foss *et al.* proposed the method KAMILA (Kay-means for mixed large data sets) which can directly deal with different types of attributes and require less parameters [27]. Chen and He presented a self-adaptive peak density clustering approach ACC-FSFDP by employing the idea of density clustering [28]. Ji *et al* introduced CCS-K-Prototypes by utilizing cuckoo search strategy and the k-prototypes framework [4].

The majority of the existing clustering algorithms are single-view learning. Multi-view learning approaches, which utilize the consistency and complementary information in different views, demonstrates more effective, more promising, and better generalization ability than the single-view counterparts in many problems [29]. The existing multi-view learning algorithms fall into one of the following three types: co-training, multiple kernel learning, and subspace learning [29]. In addition, most of the multi-view learning approaches are designed for supervised learning tasks.

With the success of multi-view learning, multi-view clustering has attracted more and more attention in recent years. Kailing *et al* introduced a multi-view density-based clustering approach, which is the multi-view version of the well-established DBSCAN [30]. Bickel and Scheffer proposed a multi-view clustering algorithm for text data and demonstrated that the multi-view spherical k-means and the multi-view EM algorithms achieved better performance than their single-view counterparts [31]. Chaudhuri *et al* presented a multi-view clustering algorithm on the basis of canonical correlation analysis [32]. Kumar *et al* introduced a co-regularized multi-view spectral clustering algorithm [33]. Wang *et al* proposed a group detection framework on the basis of multi-view clustering [34]. Huang *et al* presented a deep multi-view spectral clustering algorithm MVSCN [3]. Li *et al* introduced a multi-view spectral clustering algorithm by utilizing bipartite graphs [35]. Zhao, Ding, and Fu introduced a deep matrix factorization model for multi-view clustering on the basis of semi-nonnegative matrix factorization [36]. Nie, Li, and Li proposed a self-weighted multi-view clustering algorithm by utilizing multiple graphs [37]. Zhang *et al* presented a binary multi-view clustering algorithm BMVC to deal with image data [38].

These algorithms are either single view ones or not designed for data with both numeric and categorical attributes. In this article we present a novel multi-view clustering algorithm based on the k-prototypes framework (Multi-view K-Prototypes) for mixed numeric and categorical data. To the best of our knowledge, our algorithm is the first multi-view version of the well-known k-protypes algorithm. In our approach, we first present the representation prototype and updating approaches for cluster centres in the scenario of multiple views. Then we develop the representation of consensus prototypes and the approach to output the final clustering result, and we use a simple example to illustrate the work process of the presented Multi-view K-Prototypes algorithm. Finally, we present the complexity analysis of the Multi-view K-Prototypes algorithm and assess the performance of this algorithm on several benchmark datasets.

The remainder of this article is organized as follows: we first review some related work in Section II. In Section III, we depict the proposed approach. This is followed by the experimental results which demonstrate the effectiveness of the proposed approach in Section IV. Finally, we draw conclusions of this article and explore the future directions in Section V.

II. NOTATIONS AND RELATED METHODS

In this section, we first introduce the notations utilized in this article, and we then briefly review the idea of the multi-view EM and the k-prototypes algorithm.

A. NOTATIONS

Suppose $X = \{x_1, x_2, \ldots, x_n\}$ represents a dataset with n data objects and $x_i\,(1 \le i \le n)$ denotes a data object characterized by m attributes $A_1, A_2, \ldots, A_m$. All values of an attribute A_j in a dataset constitute the domain of values $Dom\left(A_j\right)$. For the mixed data, the domain of values could be classified into two types: numeric and categorical. The numeric domain is consisted of continuous real numbers, whereas the categorical domain is a finite set of the categorical values without natural ordering such as red, white, blue. The categorical domain is generally expressed as $Dom\left(A_j\right) = \left\{a_j^1, a_j^2, \ldots a_j^t\right\}$, where the superscript t represents the number of values of the categorical attribute A_j in a dataset. A data object x_i is usually expressed as $[A_1 = x_{i1}] \bigwedge [A_2 = x_{i2}] \bigwedge \ \ldots \ \bigwedge \left[A_j = x_{ij}\right] \bigwedge \ \ldots \ \bigwedge [A_m = x_{im}]$, where $x_{ij} \in Dom\left(A_j\right)$ for $1 \le j \le m$. For ease of description, x_i is denoted as a vector $[x_{i1}, x_{i2}, \ldots, x_{im}]$.

B. MULTI-VIEW EM

The multi-view EM clustering framework was introduced by Bickel and Scheffer for document clustering [31]. This

algorithm is a co-training style multi-view learning algorithm where the available attributes of data objects are divided into two distinct views. The multi-view EM approach, as a co-training style algorithm, has two assumptions: a) sufficiency: each view suffices for learning; b) conditional independence: the two views are conditionally independent given a mixture component [29], [31]. Let V^1 and V^2 be the two views of attributes, then a data object x_i is expressed as $\left[x_i^1, x_i^2\right]$. Here, x_i^1 and x_i^2 are vectors over the views V^1 and V^2, respectively. The process of the multi-view EM approach is presented in the Algorithm 1.

Algorithm 1 Multi-view EM

Input: A dataset $X = \left\{\left[x_1^1, x_1^2\right], \ldots, \left[x_n^1, x_n^2\right]\right\}$.

1. Initialize model parameters Θ_0^2, the maximum number of iterations T, and set the iteration number $t = 0$.
2. E step in View 2: calculate expectation for hidden variables based on the model parameters Θ_0^2.
3. Do until the stop condition is achieved:
   a) For v = 1,2:
      i. t = t + 1;
      ii. M step in view v: search for the model parameters Θ_t^v which maximize the likelihood for the data based on the expected values of the hidden variables in view $\bar{v}$ of iteration *t-1*;
      iii. E step in view v: calculate expectation for hidden variables based on the model parameters Θ_t^v;
   b) End For.
4. **Output**: the combined model parameters $\hat{\Theta} = \Theta_{t-1}^1 \cup \Theta_t^2$.

In Step 3.a.ii, View $\bar{v}$ is the complementary view of View v, and the stop condition of the multi-view EM is as follows: the iteration number t is no less than the maximum number of iterations T.

C. THE K-PROTOTYPES ALGORITHM

As mentioned above, Huang presented the well-known k-prototypes algorithm for clustering mixed data [18]. The objective of this algorithm is to divide the dataset X into k clusters or groups by minimizing the cost function, which is given as follows:

$$E(U, V) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} d\left(x_i, V_j\right), \tag{1}$$

where V_j is the cluster centre or prototype of a cluster j; $u_{ij} (0 \leq u_{ij} \leq 1)$ is an element of the membership matrix $U_{n\times k}$; and $d\left(x_i, V_j\right)$ is the distance measure which is formulated by:

$$d\left(x_i, V_j\right) \sum_{l=1}^{m} d\left(x_{il}, v_{jl}\right). \tag{2}$$

The term $d\left(x_{il}, v_{jl}\right)$ in Equation (2) is given as follows:

$$d\left(x_{il}, v_{jl}\right) = \begin{cases} \left(x_{il} - v_{jl}\right)^2 & \text{if the } l\text{th attribute is numeric,} \\ \beta_j \theta\left(x_{il}, v_{jl}\right) & \text{if the } l\text{th attribute is categorical,} \end{cases} \tag{3}$$

where $\theta(a, b) = 1$ if terms a and b have different values,$\theta(a, b) = 0$ if terms a and b have the same value, and β_j is the weight of categorical attributes in a cluster j. When the lth attribute is the numeric one, v_{jl} is the mean of the lth attribute in Cluster j; when the lth attribute is the categorical one, v_{jl} is the most frequent value or the mode of the lth attribute in Cluster j. The process of the k-prototypes algorithm is given in Algorithm 2.

Algorithm 2 K-prototypes

Input: Dataset X, the number of clusters k, β_j.

1. Choose k data objects in a random manner from the dataset X as the initial cluster centres or prototypes.
2. For each data object in X, assign it to the cluster whose prototype is the nearest one to this data object according to Equation (2); following each assignment, update the cluster centre or the prototype of the corresponding cluster.
3. Recalculate the similarity between data objects and the prototypes after all data objects have been assigned. If the nearest prototype of a data object belongs to another cluster, remove this data object from its current cluster and reassign it to the nearest one. Update the prototypes for these two clusters.
4. Repeat Step 3 until no data object changes its clusters.
5. **Output**: the clustering result.

III. THE PROPOSED METHOD

In this section, we first develop the representation and updating approaches of cluster centres in the multi-view scenario, and we then present the concept of consensus prototype to output the final clustering result. Then we depict the Multi-view K-Prototypes (multi-view clustering based on k-prototypes) approach. Finally, we use a simple example to illustrate the work process of the Multi-view K-Prototypes algorithm and analyse the complexity of this algorithm.

A. THE REPRESENTATION OF CLUSTER CENTRES

In this subsection, we present a representation prototype of the cluster centre in the multi-view scenario. Like the other co-training style multi-view learning approaches, we split the available attributes of data objects into two views. As aforementioned, a data object is described by m attributes. For ease of description, let the first u attributes are in View 1, and the rest attributes are in View 2. Thus, in the multi-view scenario, a data object x_i is represented as

$$x_i = \left[x_{i,1}^1, x_{i,2}^1, \ldots, x_{i,u}^1, x_{i,u+1}^2, x_{i,u+2}^2, \ldots, x_{i,m}^2\right]. \tag{4}$$

In (4), x_i can be abbreviated as $x_i = \left[x_i^1, x_i^2\right]$. The representation prototype of the cluster centre for a cluster j is expressed as

$$v_j = \left[v_{j,1}^1, v_{j,2}^1, \ldots, v_{j,u}^1, v_{j,u+1}^2, v_{j,u+2}^2, \ldots, v_{j,m}^2\right]. \quad (5)$$

In (5), v_j can be abbreviated as $v_j = \left[v_j^1, v_j^2\right]$. In a cluster centre v_j, if the lth attribute in View e is a numeric attribute, $v_{j,l}^e$ is the mean of that attribute in Cluster j; if the lth attribute in View e is a categorical one, $v_{j,l}^e$ is the mode or the most frequent value of that attribute in Cluster j.

B. THE MULTI-VIEW K-PROTOTYPES ALGORITHM

In this subsection, we propose a novel multi-view clustering algorithm called Multi-view K-Prototypes for dealing with mixed numeric and categorical data. Like the other co-training style multi-view learning algorithms, we split the available attributes of data objects into two views. There is a cost function in each view. The goal of the Multi-view K-Prototypes algorithm is to divide a dataset X into k clusters by minimizing the cost function in each view. In a view e, the cost function is given as follows:

$$E\left(U^e, V^e\right) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^e d(x_i^e, v_j^e), \quad (6)$$

where U^e is the membership matrix over View e;$V^e = \left[v_1^e, v_2^e, \ldots, v_k^e\right]$ is the set of cluster centre prototypes over View e; x_i^e is the data object x_i over View e; v_j^e is the cluster centre prototype of Cluster j over View e, $u_{ij}^e\left(0 \le u_{ij}^e \le 1\right)$ is an element of the membership matrix $U_{n\times k}^e$, and $d\left(x_i^e, v_j^e\right)$ is the dissimilarity measure which is given as follows:

$$d\left(x_i^e, v_j^e\right) = \sum_{l=1,e'=e}^{m} d\left(x_{il}^{e'}, v_{jl}^{e'}\right). \quad (7)$$

In (7), e' denote the view where the lth attribute is located in, and $d\left(x_{il}^{e'}, v_{jl}^{e'}\right)$ is formulated as follows:

$$d\left(x_{il}^{e'}, v_{jl}^{e'}\right) = \begin{cases} \left(\dfrac{x_{il}^{e'} - v_{jl}^{e'}}{max_l - min_l}\right)^2 & \text{if the } l\text{th attribute is numeric,} \\ \beta_j \theta\left(x_{il}^{e'}, v_{jl}^{e'}\right) & \text{if the } l\text{th attribute is categorical,} \end{cases} \quad (8)$$

where $x_{il}^{e'}$ represents the value of the lth attribute of a data object x_i over View e'; max_l and min_l are the maximum and minimum value of the lth attribute in Dataset X, respectively; β_j is the weight of the categorical attributes in Cluster j; the value of $\theta\ (a, b)$ is 0 if the terms a and b have the same value; the value of $\theta\ (a, b)$ is 1 if the terms a and b have different values. When the lth attribute is a numeric one, $v_{jl}^{e'}$ is the mean of the lth attribute in Cluster j over the view e'; when the lth attribute is a categorical one, $v_{jl}^{e'}$ is the mode or the most frequent value of the lth categorical attribute in the cluster j over the view e'. Let v_j^e be a cluster centre over View e, Cluster c_j^e is given as follows:

$$c_j^e = \left\{x_i^e \in X^e : d\left(x_i^e, v_j^e\right) < d\left(x_i^e, v_b^e\right), j \ne b\right\}, \quad (9)$$

where X^e denotes the dataset X over View e. In (6), u_{ij}^e is given as follows:

$$u_{ij}^e = \begin{cases} 1, & if\ x_i^e \in c_j^e, \\ 0, & if\ x_i^e \notin c_j^e. \end{cases} \quad (10)$$

Based on the above descriptions, we present the flow chart of the proposed Multi-view K-Prototypes approach in Figure 1, and depict the detailed process of this approach in Algorithm 3.

Algorithm 3 Multi-view K-Prototypes

Input: Dataset $X = \left\{\left[x_1^1, x_1^2\right], \left[x_2^1, x_2^2\right], \ldots, \left[x_n^1, x_n^2\right]\right\}$, the maximization iteration number *MaxIteN*, the fixed iteration number *FixedIteN*, and the expected number of clusters k.

1. Initialize the cluster centre prototype V_t^2 over View 2 randomly, and set the iteration number $t = 0$;
2. Partition data objects over View 2: based on the cluster centre prototype V_t^2, allocate each data object in the dataset X to the cluster whose cluster centre is nearest to this data object according to (9) over View 2;
3. **While (the stop criterion is not met)**
   a) **For e=1, 2:**
      i. ***t=t+1;***
      ii. Update the cluster centre prototypes over View e: calculate the cluster centre prototypes $V_t^e = \left\{v_1^e, v_2^e, \ldots, v_k^e\right\}$ based on the partition over View $\bar{e}$ of the iteration t-1;
      iii. Partition data objects over View e: based on the cluster centre prototypes V_t^e, allocate each data object in the dataset X to the cluster whose cluster centre is nearest to this data object according to (9) over View e;
   b) **End For**;
4. **End While**
5. Return combined cluster centre prototypes $ComV = V_{t-1}^1 \cup V_t^2$, and output the final clustering result.

In Step 3.a.ii of Algorithm 3, View $\bar{e}$ is the complementary view of View e, and the updating process of cluster centres (prototypes) is described as follows:

For a cluster centre prototype v_j^e, if the lth attribute is a numeric one, v_{jl}^e is calculated according to (11).

$$v_{jl}^e = \frac{\sum_{i=1, x_i^{\bar{e}} \in c_j^{\bar{e}}}^{n} x_{il}^e}{\left|c_j^{\bar{e}}\right|}, \quad (11)$$

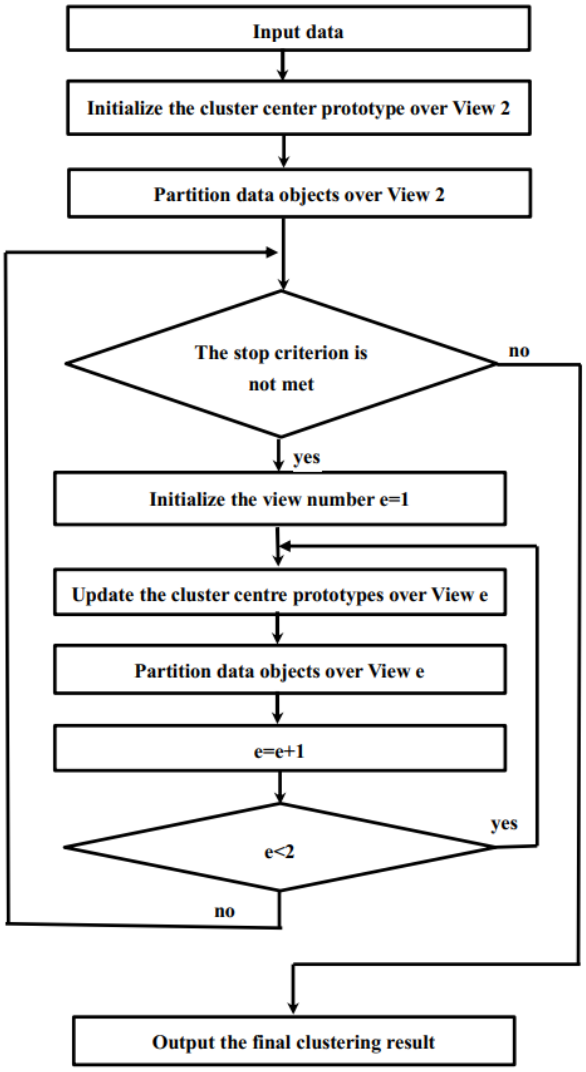

FIGURE 1. The flow chart of the Multi-view K-Prototypes algorithm.

where $c_j^{\bar{e}}$ is the cluster j over View $\bar{e}$, and the symbol $\left|c_j^{\bar{e}}\right|$ |.| denotes the number of data objects in Cluster $c_j^{\bar{e}}$. If the lth attribute is a categorical one, v_{jl}^e is the most frequent value or the mode of the lth attribute over View e for the data objects in the cluster $c_j^{\bar{e}}$, which is given as follows:

$$v_{jl}^e = calculate\ Mode(Vals), \tag{12}$$

where *Vals* is formulated as follows:

$$Vals = \{x_{il}^e : x_i^{\bar{e}} \in c_j^{\bar{e}}. \tag{13}$$

Once an iteration is completed, we calculate the cost function for each view e according to (6). If the value of the cost function is not improved for a given number of iterations such as 5 in each view, we terminate the optimization process. Therefore, the stop criteria in our Multi-view K-Prototypes algorithm are summarized as follows: the iteration number t is no less than the maximum iteration number (*MaxIteN*) or the cost functions are not improved for a fixed number of iterations (*FixedIteN*) in each view.

When the iteration process of the Multi-view K-Prototypes algorithm ends, the data objects in Cluster c_j^1 and the ones in Cluster c_j^2 may not be identical. Inspired by Bickel and Scheffer's work, we also utilize the consensus principle, which aims to maximize the agreement on multiple views, to gain the final clustering result [29], [31]. Firstly, we calculate the consensus partition, which is given by:

$$CP = \{cp_1, cp_2, \ldots, cp_k\}, \tag{14}$$

where cp_j is formulated as follows:

$$cp_j = \left\{x_i \in X : x_i^1 \in c_j^1 \bigwedge x_i^2 \in c_j^2\right\}. \tag{15}$$

Based on the consensus partition *CP*, we then calculate the consensus cluster centre prototypes, which are given by:

$$CV = \{cv_1, cv_2, \ldots, cv_k\}. \tag{16}$$

Like the representation of a data object x_i, a consensus cluster centre prototype cv_j is expressed as $cv_j = \left[cv_j^1, cv_j^2\right]$. In a consensus cluster centre prototype cv_j, if the lth attribute is a numeric one, the cv_{jl}^e is calculated according to (17) as follows:

$$cv_{jl}^e = \frac{\sum_{i=1, x_i \in cp_j}^{n} x_{il}^e}{\left|cp_j\right|}, \tag{17}$$

where the symbol $\left|cp_j\right|$ |.| denotes the number of data objects in the cluster cp_j; if the lth attribute is a categorical one, the cv_{jl}^e is the mode or the most frequent value of the lth attribute in the cluster cp_j, which can be formulated as follows:

$$cv_{jl}^e = calculate\ Mode\ (CVals), \tag{18}$$

where *CVals* is defined as follows:

$$CVals = \left\{x_{il}^e : x_i \in cp_j\right\}. \tag{19}$$

Finally, we allocate each data object in a dataset X to the cluster with the nearest consensus cluster centre as follows:

$$c_j = \left\{x_i \in X : d\left(x_i, cv_j\right) < d\left(x_i, cv_s\right), s \neq j\right\} \tag{20}$$

In (20), $d\left(x_i, cv_j\right)$ is given as follows:

$$d\left(x_i, cv_j\right) = \sum_{e=1}^{2} d\left(x_i^e, cv_j^e\right). \tag{21}$$

C. AN ILLUSTRATIVE EXAMPLE

In this subsection, we first give a simple synthetic dataset and then utilize this dataset to illustrate the execution process of the Multi-view K-Prototypes algorithm. The synthetic dataset X consists of eight data objects where each one has two numeric attributes and two categorical attributes. In Table 1, we depict the details of these data objects.

Assume that the synthetic dataset X has two clusters, the attributes in View 1 are gender and age, and the attributes in View 2 are height and hobby. Let the clusters number k be 2, the maximization iteration number *MaxIteN* be 4,

TABLE 1. The Synthetic Dataset

ID \ Attributes	Gender	Age	Height(cm)	hobby
x_1	male	19	176	music
x_2	female	23	166	writing
x_3	male	36	185	football
x_4	female	24	175	tennis
x_5	male	32	194	basketball
x_6	male	26	182	tennis
x_7	female	22	170	writing
x_8	male	25	174	music

and the fixed iteration number *FixedIteN* be 2. As in the original K-Prototypes algorithm, the parameter β_j is set as 1.0. In the stage of initialization, the cluster centre prototype V_t^2 is initialized as two randomly selected data objects over View 2, and the current iteration number t is set as 0. Assume that data objects x_2 and x_5 are chosen, the cluster centre prototype V_0^2 are listed as follows:

$$V_0^2 = \{166, \textit{writing}; 194, \textit{basketball}\}.$$

Based on the cluster centre prototype V_0^2, the data objects in the dataset X are divided into two clusters according to (9). These two clusters are listed as follows:

$$c_1^2 = \left\{x_1^2, x_2^2, x_4^2, x_7^2, x_8^2\right\},$$
$$c_2^2 = \left\{x_3^2, x_5^2, x_6^2\right\}.$$

The value of the cost function View 2, which is calculated according to (6), is 5.62. The iteration number t increases to 1. Based on the partition over View 2, the representation prototype V_1^1 of cluster centre over View 1 is calculated based on (11), (12) and (13). The representation prototype V_1^1 is listed as follows:

$$V_1^1 = \{\textit{female}, 22.6; \textit{male}31.33\}.$$

Based on the cluster centre prototype V_1^1, the data objects in the dataset X are divided into two clusters according to (9). These two clusters are listed as follows:

$$c_1^1 = \left\{x_2^1, x_4^1, x_7^1\right\}, c_2^1 = \left\{x_1^1, x_3^1, x_5^1, x_6^1, x_8^1\right\}.$$

The value of the cost function over View 1, which is calculated by (6), is 0.85. The iteration number t increases to 2. Based on the partition over View 1, the representation prototype V_2^2 of cluster centre over View 2 is calculated by (11), (12) and (13). The cluster centre prototype V_2^2 is listed as follows:

$$V_2^2 = \{170.33, \textit{writing}; 182.2, \textit{music}\}.$$

Based on the cluster centre prototype V_2^2, the data objects in the dataset X are divided into two clusters by using (9). These two clusters are listed as follows:

$$c_1^2 = \left\{x_2^2, x_4^2, x_7^2\right\}, c_2^2 = \left\{x_1^2, x_3^2, x_5^2, x_6^2, x_8^2\right\}.$$

The value of the cost function over View 2, which is calculated by (6), is 4.37. This value is smaller than the previous value 5.62 over View 2, which means the value of the objective cost function over View 2 is improved or reaches a new minimum. The iteration number t increases to 3. Based on the partition over View 2, the representation prototype V_3^1 of cluster centre over View 1 is calculated by (11), (12) and (13). The cluster centre prototype V_3^1 is listed as follows:

$$V_3^1 = \{\textit{female}, 23.0; \textit{male}, 27.6\}.$$

Based on the cluster centre prototype V_3^1, the data objects in the dataset X are divided into two clusters by using (9). These two clusters are listed as follows:

$$c_1^1 = \left\{x_2^1, x_4^1, x_7^1\right\}, c_2^1 = \left\{x_1^1, x_3^1, x_5^1, x_6^1, x_8^1\right\}.$$

The value of the cost function over the view 1, which is calculated by (6), is 0.61. This value is smaller than the previous value 0.85 over the view 1, which means the value of the cost function over the view 1 is improved or reaches a new minimum. The iteration number t increases to 4. Based on the partition over View 1, the representation prototype V_4^2 of cluster centre over View 2 is calculated by using (11), (12) and (13). The cluster centre prototype V_4^2 is listed as follows:

$$V_4^2 = \{170.33, \textit{writing}; 182.2, \textit{music}\}.$$

Based on the cluster centre prototype V_4^2, the data objects in the dataset X are divided into two clusters by using (9). These two clusters are listed as follows:

$$c_1^2 = \left\{x_2^2, x_4^2, x_7^2\right\}, c_2^2 = \left\{x_1^2, x_3^2, x_5^2, x_6^2, x_8^2\right\}.$$

The value of the cost function over the view 2, which is calculated by using (6), is 4.37. This value is equal to the previous value 4.37 over the view 2, which means the value of the cost function over View 2 is not improved or not reaches a new minimum. Due to iteration number t is no less than the maximum iteration number *MaxIteN*, the execution process of the Multi-view K-Prototypes algorithm is terminated. To obtain the final clustering result, we firstly get the consensus partition *CP* using (14) and (15). The elements in the consensus partition are listed as follows:

$$cp_1 = \left\{[x_2^1, x_2^2], [[x_4^1, x_4^2], [x_7^1, x_7^2]\right\},$$
$$cp_2 = \left\{\left[x_1^1, x_1^2\right], \left[x_3^1, x_3^2\right], \left[x_5^1, x_5^2\right], \left[x_6^1, x_6^2\right], [x_8^1, x_8^2]\right\}.$$

Based on the consensus partition *CP*, the consensus cluster centre prototypes *CV* is calculated by employing (16), (17), (18) and (19). The elements in the consensus cluster centre prototypes *CV* are listed as follows:

$$cv_1 = \{\textit{female}, 23.0, 170.33, \textit{writing}\},$$
$$cv_2 = \{\textit{male}, 27.6, 182.2, \textit{music}\}.$$

TABLE 2. The Datasets Used in Experiments

Dataset	Number of numeric attributes	Number of categorical attributes	Number of data objects	Number of classes
Zoo	1	16	101	7
Heart disease (Case 1)	6	9	303	5
Heart disease (Case 2)	6	8	303	2
Credit approval	6	10	690	2
Soybean	0	36	47	4
Breast cancer	9	2	699	2

Based on the consensus cluster centre prototypes CV, the final clustering result can be obtained by dividing the data objects in the dataset X into two clusters according to (20) and (21). Therefore, the final clustering result is as follows:

$$c_1 = \{x_2, x_4, x_7\}, c_2 = \{x_1, x_3, x_5, x_6, x_8\}.$$

D. ALGORITHM COMPLEXITY ANALYSIS

In this subsection, we discuss the complexity of the Multi-view K-Prototypes algorithm. The time complexity of the presented approach includes two main components: the updating of cluster centre prototypes, and the calculation of the partition matrix of data objects in each iteration. The computational cost of these two components are $O(k(p + Nm - Np)n)$ and $O(nkm)$, respectively. Here k denotes the number of clusters; p denotes the number of numeric attributes; m denotes the number of all attributes; N denotes the maximal number of values for all categorical attributes; n denotes the number of data objects contained in a dataset X. Thus, the whole time complexity of the Multi-view K-Prototypes algorithm is $O(k(p + Nm - Np + m)ns)$, where s denotes the number of iterations required when this algorithm terminates. As for the space complexity, the Multi-view K-Prototypes algorithm needs $O(mn)$ to store the dataset X, $O(nk)$ to store the membership matrix, and $O(km)$ to store the cluster centre prototypes. Thus, the whole space complexity of the Multi-view K-Prototypes algorithm is $O(mn+km+nk)$.

IV. EXPERIMENTS AND DISCUSSION

For evaluating the performance of the Multi-view K-Prototypes algorithm, we run this algorithm on four datasets: zoo, heart disease, credit approval, and breast cancer. All these datasets are downloaded from the well-known UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.php). In Table 2, we briefly list the important information of these datasets.

In clustering analysis, the Rand Index [39] and the Yang's accuracy measures [40] are widely utilized for evaluating the clustering results. In order to assess the gained clustering results, we employ these two measures in this research. In Yang's approach, the clustering accuracy (AC), precision (PR), and recall (RE) are respectively given by:

$$AC = \frac{\sum_{i=1}^{k} a_i}{n}, \tag{22}$$

$$PR = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i+b_i}}{k}, \tag{23}$$

$$RE = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i+e_i}}{k}. \tag{24}$$

In (22)-(24), a_i denotes the number of data objects correctly distributed to the class C_i; b_i denotes the number of data objects wrongly allocated to the class C_i; e_i denotes the number of data objects wrongly refused from the class C_i; k represents the number of classes for a dataset, and n denotes the number of data objects included in a dataset. If $Y = \{y_1, y_2, \ldots, y_{v_1}\}$ and $Y' = \{y'_1, y'_2, \ldots, y'_{v_2}\}$ are two partitions of a dataset $X = \{x_1, x_2, \ldots, x_n\}$, then the Rand Index (RI) [39] is formulated as follows:

$$RI = \frac{\sum_{i=1,j=2;i<j}^{n} \omega_{ij}}{\binom{n}{2}}, \tag{25}$$

where ω_{ij} is given by:

$$\omega_{ij} = \begin{cases} 1, & \text{if there exist } v \text{ and } v' \text{such that both } x_i \\ & \text{and } x_j \text{ are in both } y_v \text{ and } y'_{v'}, \\ 1, & \text{if there exist } v \textit{ and } v' \text{such that } x_i \text{is in both} \\ & y_v \text{ and } y_{v'} \text{' while } x_j \text{ is in neither } y_v \text{ nor } y'_{v'}, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

The higher values of these four measures (i. e. AC, PR, RE, and RI) mean the better clustering result. To evaluate the algorithm's performance, we assessed the proposed Multi-view K-Prototypes algorithm on four datasets. Owing to the randomness of the initial cluster centres, the Multi-view K-Prototypes algorithm is executed twenty trials on each dataset, and the mean values of AC, PR, RE, and RI are obtained. The clustering results of several algorithms, including K-prototypes [18], SBAC [22], KL-FCM-GM [21], EKP [25], ABC-K-Prototypes [41], CCS-K-Prototypes [4], and ACC-FSFDP [28] reported in [4] are also provided for comparison. In our Multi-view K-Prototypes algorithm, we set the maximum iteration number $MaxIteN$ as 100, the fixed iteration number $FixedIteN$ as 5 by the rule of thumb, and the number of clusters k as the number of classes contained in a dataset. In K-Prototypes algorithm [18], the weight parameter is set as 1.0 in all experiments. For a fair comparison of our proposed approach with

the K-Prototypes algorithm, the parameter β_j in Eq. (8) is also set as 1.0 in our experiments. Similar to the Bickel and Scheffer's approach [29], [31], we split the available attributes of data objects in a random way into two views to construct multiple views for the data without natural multiple views.

We start our experiments by considering the zoo dataset. This dataset includes 101 data objects, where each one has one numeric attribute and sixteen categorical attributes. According to the class attribute, the zoo dataset has seven classes. In Table 3a, we list the accuracy (AC) of Multi-view K-Prototypes and other algorithms used for comparison on the zoo dataset. From this table we can see that our Multi-view K-Prototypes algorithm achieves the highest AC value (0.899) than other algorithms. In Table 3b, we summarize the precision (PR) of the above algorithms on the zoo dataset. We can see that the Multi-view K-Prototypes algorithm achieves the PR value of 0.859, which is comparable with other algorithms. In Table 3c, we list the recall (RE) of all algorithms on the zoo dataset. The Multi-view K-Prototypes algorithm achieves the highest RE of 0.734. In Table 3d, we summarize the rand index (RI) of all algorithms on the zoo dataset. Again, our Multi-view K-Prototypes algorithm achieves the highest RI of 0.939. The clustering results in Tables 3a-3d clearly show that our proposed Multi-view K-Prototypes algorithm obtains the highest values on the measures AC, RE, RI, and achieves a comparable value on the measure PR.

The heart disease dataset consists of 303 data objects. These data objects are the patient instances with 6 numeric attributes and 9 categorical attributes. There are two class attributes in this dataset. If the 15*th* attribute is used as the class attribute, the data object is characterized by 14 attributes, and the heart disease dataset has five classes; if the 14*th* attribute is used as the class attribute, the data object is characterized by 13 attributes, and the heart disease dataset has two classes. For the first case, we summarize the accuracy (AC) of Multi-view K-Prototypes and other algorithms on the heart disease dataset (first case) in Table 4a. From this table we can see that our Multi-view K-Prototypes algorithm achieves the highest AC value (0.656) than other algorithms. In Table 4b, we list the precision (PR) of the above algorithms on the heart disease dataset (first case). We can see that the Multi-view K-Prototypes algorithm achieves the PR value of 0.637, which is comparable with other algorithms. In Table 4c, we summarize the recall (RE) of all algorithms on the heart disease dataset (first case). The Multi-view K-Prototypes algorithm achieves the highest RE of 0.398. In Table 4d, we list the rand index (RI) of all algorithms on the heart disease dataset (first case). Again, our Multi-view K-Prototypes algorithm achieves the highest RI of 0.684. The clustering results in Tables 4a-4d clearly show that our proposed Multi-view K-Prototypes algorithm obtains the highest values on the measures AC, RE, and RI, and achieves a comparable value on the measure PR.

TABLE 3. a. The Accuracy (AC) of the Clustering Algorithms on the Zoo Dataset b.The precision (PR) of the clustering algorithms on the Zoo dataset. c. The recall (RE) of the clustering algorithms on the Zoo dataset. d. The rand index (RI) of the clustering algorithms on the Zoo dataset.

(a)

Algorithms	AC
K-Prototypes	0.806
SBAC	0.426
KL-FCM-GM	0.870 ($\alpha = 1.3$)
EKP	0.628
ABC-K-Prototypes	0.886
ACC-FSFDP	0.874
CCS-K-Prototypes	0.888 (N=40, pa=0.2)
Multi-view K-Prototypes	**0.899**

(b)

Algorithms	PR
K-Prototypes	0.827
SBAC	0.484
KL-FCM-GM	0.844 ($\alpha = 1.3$)
EKP	0.729
ABC-K-Prototypes	0.861
ACC-FSFDP	0.862
CCS-K-Prototypes	**0.873** (N=30, pa=0.2)
Multi-view K-Prototypes	0.859

(c)

Algorithms	RE
K-Prototypes	0.636
SBAC	0.172
KL-FCM-GM	0.685 ($\alpha = 1.3$)
EKP	0.419
ABC-K-Prototypes	0.718
CCS-K-Prototypes	0.709 (N=30, pa=0.3)
Multi-view K-Prototypes	**0.734**

(d)

Algorithms	RI
K-Prototypes	0.857
SBAC	0.648
KL-FCM-GM	0.918 ($\alpha = 1.8$)
EKP	0.601
ABC-K-Prototypes	0.894
CCS-K-Prototypes	0.901 (N=20, pa=0.25)
Multi-view K-Prototypes	**0.939**

For the second case where the heart disease dataset uses the 14th attribute as its class attribute, and has two classes. In Table 5a, we list the accuracy (AC) of Multi-view K-Prototypes and other algorithms used for comparison on the heart disease dataset (second case). From this table we can see that our Multi-view K-Prototypes algorithm obtains the AC value of 0.810, which is comparable with other algorithms. In Table 5b, we summarize the precision (PR) of the above algorithms on the heart disease dataset (second case).

TABLE 4. a. The Accuracy (AC) of the Clustering Algorithms on the Heart Disease Dataset (First Case) b. The precision *(PR)* of the clustering algorithms on the Heart disease dataset (first case). c. The recall *(RE)* of the clustering algorithms on the Heart disease dataset (first case). d. The rand index (RI) of the clustering algorithms on the Heart disease dataset (first case).

(a)

Algorithms	AC
K-Prototypes	0.547
SBAC	0.545
KL-FCM-GM	0.653 ($\alpha = 1.2$)
EKP	0.545
ABC-K-Prototypes	0.648
CCS-K-Prototypes	0.648 (N=30, pa=0.2)
Multi-view K-Prototypes	**0.656**

(b)

Algorithms	PR
K-Prototypes	0.521
SBAC	0.566
KL-FCM-GM	**0.766** ($\alpha = 1.9$)
EKP	0.109
ABC-K-Prototypes	0.658
CCS-K-Prototypes	0.675 (N=40, pa=0.15)
Multiview-K-Prototypes	0.637

(c)

Algorithms	RE
K-Prototypes	0.216
SBAC	0.2
KL-FCM-GM	0.395 ($\alpha = 1.4$)
EKP	0.2
ABC-K-Prototypes	0.379
CCS-K-Prototypes	0.388 (N=35, pa=0.15)
Multi-view K-Prototypes	**0.398**

(d)

Algorithms	RI
K-Prototypes	0.601
SBAC	0.503
KL-FCM-GM	0.673 ($\alpha = 1.2$)
EKP	0.355
ABC-K-Prototypes	0.667
CCS-K-Prototypes	0.680 (N=35, pa=0.25)
Multi-view K-Prototypes	**0.684**

TABLE 5. a. The accuracy (AC) of the Clustering Algorithms on the Heart Disease Dataset (Second Case) b. The precision (PR) of the clustering algorithms on the Heart disease dataset (second case). c. The recall (RE) of the clustering algorithms on the Heart disease dataset (second case). d. The rand index (RI) of the clustering algorithms on the Heart disease dataset (second case).

(a)

Algorithms	AC
K-Prototypes	0.577
SBAC	0.545
KL-FCM-GM	0.762 ($\alpha = 1.7$)
EKP	0.545
ABC-K-Prototypes	0.809
CCS-K-Prototypes	**0.812** (N=20, pa=0.3)
Multi-view K-Prototypes	0.810

(b)

Algorithms	PR
K-Prototypes	0.570
SBAC	0.567
KL-FCM-GM	0.783 ($\alpha = 2.6$)
EKP	0.272
ABC-K-Prototypes	0.808
CCS-K-Prototypes	**0.812** (N=20, pa=0.3)
Multi-view K-Prototypes	0.809

(c)

Algorithms	RE
K-Prototypes	0.566
SBAC	0.5
KL-FCM-GM	0.768 ($\alpha = 1.7$)
EKP	0.5
ABC-K-Prototypes	0.806
CCS-K-Prototypes	**0.809** (N=20, pa=0.3)
Multi-view K-Prototypes	0.807

(d)

Algorithms	RI
K-Prototypes	0.510
SBAC	0.499
KL-FCM-GM	0.641 ($\alpha = 1.7$)
EKP	0.502
ABC-K-Prototypes	0.689
CCS-K-Prototypes	**0.694** (N=20, pa=0.3)
Multi-view K-Prototypes	0.691

We can see that the Multi-view K-Prototypes algorithm achieves the PR value of 0.809, which is comparable with other algorithms. In Table 5c, we list the recall (RE) of all algorithms on the heart disease dataset (second case). The Multi-view K-Prototypes algorithm achieves the RE value of 0.807, which is comparable with other algorithms. In Table 5d, we summarize the rand index (RI) of all algorithms on the heart disease dataset (second case). Again, our Multi-view K-Prototypes algorithm achieves the RI value of 0.691, which is comparable with other algorithms. The clustering results in Tables 5a-5d show that our proposed Multi-view K-Prototypes algorithm obtains the comparable values on the measures AC, PR, RE, and RI.

The credit approval dataset consists of 690 customer instances derived from credit card organizations. The data objects in this dataset are described by six numeric attributes and ten categorical attributes. According to the class attribute, the credit approval dataset has two classes. In Table 6a, we list the accuracy (AC) of Multi-view K-Prototypes and other algorithms used for comparison on the credit approval dataset. From this table we can see that our Multi-view

TABLE 6. a. The Accuracy (AC) of the Clustering Algorithms on the Credit Approval Dataset b. The precision (PR) of the clustering algorithms on the Credit approval dataset. c. The recall (RE) of the clustering algorithms on the Credit approval dataset. d. The rand index (RI) of the clustering algorithms on the Credit approval dataset.

(a)

Algorithms	AC
K-Prototypes	0.562
SBAC	0.555
KL-FCM-GM	0.578 ($\alpha = 2.4$)
EKP	0.686
ABC-K-Prototypes	0.794
ACC-FSFDP	0.784
CCS-K-Prototypes	0.796 (N=30, pa=0.2)
Multi-view K-Prototypes	**0.812**

(b)

Algorithms	PR
K-Prototypes	0.780
SBAC	0.558
KL-FCM-GM	0.642 ($\alpha = 2.4$)
EKP	0.724
ABC-K-Prototypes	0.792
ACC-FSFDP	**0.814**
CCS-K-Prototypes	0.794 (N=30, pa=0.2)
Multi-view K-Prototypes	0.810

(c)

Algorithms	RE
K-Prototypes	0.508
SBAC	0.5
KL-FCM-GM	0.549 ($\alpha = 2.4$)
EKP	0.657
ABC-K-Prototypes	0.795
CCS-K-Prototypes	0.796 (N=30, pa=0.2)
Multi-view K-Prototypes	**0.810**

(d)

Algorithms	RI
K-Prototypes	0.507
SBAC	0.499
KL-FCM-GM	0.513 ($\alpha = 2.4$)
EKP	0.568
ABC-K-Prototypes	0.673
CCS-K-Prototypes	0.674 (N=30, pa=0.2)
Multi-view K-Prototypes	**0.695**

TABLE 7. a. The Accuracy (AC) of the Clustering Algorithms on the Breast Cancer Dataset b. The precision (PR) of the clustering algorithms on the Breast cancer dataset. c. The recall (RE) of the clustering algorithms on the Breast cancer dataset. d. The rand index (RI) of the clustering algorithms on the Breast cancer dataset.

(a)

Algorithms	AC
K-Prototypes	**0.961**
SBAC	0.655
KL-FCM-GM	0.804 ($\alpha = 1.1$)
EKP	0.701
ABC-K-Prototypes	0.959
ACC-FSFDP	0.938
CCS-K-Prototypes	0.958 (N=40, pa=0.3)
Multi-view K-Prototypes	0.956

(b)

Algorithms	PR
K-Prototypes	**0.959**
SBAC	0.650
KL-FCM-GM	0.813 ($\alpha = 1.1$)
EKP	0.767
ABC-K-Prototypes	0.958
ACC-FSFDP	0.947
CCS-K-Prototypes	0.957 (N=40, pa=0.3)
Multi-view K-Prototypes	0.955

(c)

Algorithms	RE
K-Prototypes	**0.954**
SBAC	0.500
KL-FCM-GM	0.753 ($\alpha = 1.1$)
EKP	0.771
ABC-K-Prototypes	0.952
CCS-K-Prototypes	0.949 (N=40, pa=0.3)
Multi-view K-Prototypes	0.947

(d)

Algorithms	RI
K-Prototypes	**0.925**
SBAC	0.511
KL-FCM-GM	0.686 ($\alpha = 1.1$)
EKP	0.580
ABC-K-Prototypes	0.922
CCS-K-Prototypes	0.919 (N=40, pa=0.3)
Multi-view K-Prototypes	0.915

K-Prototypes algorithm obtains the highest AC value (0.812) than other algorithms. In Table 6b, we summarize the precision (PR) of the above algorithms on the credit approval dataset. We can see that the Multi-view K-Prototypes algorithm achieves the PR value of 0.810, which is comparable with other algorithms. In Table 6c, we list the recall (RE) of all algorithms on the credit approval dataset. The Multi-view K-Prototypes algorithm achieves the highest RE value of 0.810. In Table 6d, we summarize the rand index (RI) of all algorithms on the credit approval dataset. Again, our Multi-view K-Prototypes algorithm achieves the highest RI of 0.695. The clustering results in Tables 6a-6d clearly illustrate that our proposed Multi-view K-Prototypes algorithm obtains the highest values on the measures AC, RE, RI, and achieves a comparable value on the measure PR.

The breast cancer dataset consists of 699 data objects, each of which has eleven attributes. The first attribute, as the code number of samples, is not employed in clustering process. According to the class attribute, the breast cancer dataset

has two classes. In Table 7a, we summarize the accuracy (AC) of Multi-view K-Prototypes and other algorithms on the breast cancer dataset. From this table we can see that our Multi-view K-Prototypes algorithm achieves the AC value of 0.956, which is comparable with other algorithms. In Table 7b, we list the precision (PR) of the above algorithms on the breast cancer dataset. We can see that the Multi-view K-Prototypes algorithm achieves the PR value of 0.955, which is comparable with other algorithms. In Table 7c, we summarize the recall (RE) of all algorithms on the breast cancer dataset. The Multi-view K-Prototypes algorithm achieves the RE value of 0.947, which is comparable with other algorithms. In Table 7d, we list the rand index (RI) of all algorithms on the breast cancer dataset. Again, our Multi-view K-Prototypes algorithm achieves the RI value of 0.915, which is comparable with other algorithms. The clustering results in Tables 7a-7d illustrate that our proposed Multi-view K-Prototypes algorithm obtains the comparable values on the measures AC, PR, RE, and RI.

The clustering results in Tables 3a-7d illustrate that the proposed Multi-view K-Prototypes approach obtains better results than other seven clustering algorithms in most cases. These results clearly demonstrate that our Multi-view K-Prototypes algorithm is suitable for dealing with mixed numeric and categorical data. We believe the reasons for the success of the proposed Multi-view K-Prototypes approach are as follows:

Firstly, we specifically design the representation prototype of cluster centres for the clusters with both numeric and categorical attributes in the scenario of multiple views.

Secondly, we propose the updating approaches for the cluster centres with both numeric and categorical attributes in the scenario of multiple views.

Thirdly, we design the cost function for clustering the mixed numeric and categorical data in the scenario of multiple views.

Based on the above features, the clustering process of the proposed Multi-view K-Prototypes algorithm can not only effectively deal with different types of attributes, but also utilize the complementary and diverse information in different views.

Therefore, the Multi-view K-Prototypes approach achieves superior results in most cases.

V. CONCLUSION AND FUTURE WORK

In this research, we have presented a novel multi-view clustering algorithm Multi-view K-Prototypes, which to the best of our knowledge is the first multi-view version of k-prototypes algorithm for clustering data with both numeric and categorical attributes. In our approach, we propose representation prototype and updating approaches for the cluster centres under the scenario of multiple views, design the cost function for the mixed data over different views, and develop the approach to obtain the final clustering result by integrating the clustering results on each view. These are the major contributions in this research.

Then we used a simple example to illustrate the work process of the Multi-view K-Prototypes algorithm. Finally, we tested the Multi-view K-Prototypes algorithm on four datasets in terms of the clustering accuracy (AC), precision (PR), recall (RE), and rand index (RI). The experiments results validate the excellent performance of the Multi-view K-Prototypes algorithm.

As mentioned in Section I, multiple kernel learning is one of the three parts in the multi-view learning. However, there are few works on the task of clustering mixed data. Therefore, in our future work, we will investigate the potential of multiple kernel learning on clustering mixed data.

REFERENCES

[1] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the *k*-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, Jan. 2013.

[2] A. K. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[3] Z. Huang, J. T. Zhou, X. Peng, C. Zhang, H. Zhu, and J. Lv, "Multi-view spectral clustering network," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 2563–2569.

[4] J. Ji, W. Pang, Z. Li, F. He, G. Feng, and X. Zhao, "Clustering mixed numeric and categorical data with cuckoo search," *IEEE Access*, vol. 8, pp. 30988–31003, 2020.

[5] G. Bordogna and G. Pasi, "A quality driven hierarchical data divisive soft clustering for information retrieval," *Knowl.-Based Syst.*, vol. 26, pp. 9–19, Feb. 2012.

[6] F. Naouar, L. Hlaoua, and M. N. Omri, "Collaborative information retrieval model based on fuzzy clustering," in *Proc. Int. Conf. High Perform. Comput. Simulation (HPCS)*, Genoa, Italy, Jul. 2017, pp. 495–502.

[7] Y. Xin, Z.-Q. Xie, and J. Yang, "The privacy preserving method for dynamic trajectory releasing based on adaptive clustering," *Inf. Sci.*, vol. 378, pp. 131–143, Feb. 2017.

[8] C. Luo, W. Pang, and Z. Wang, "Semi-supervised clustering on heterogeneous information networks," in *Proc. 18th Pacific–Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, Tainan, Taiwan, 2014, pp. 548–559.

[9] K. K. Bharti and P. K. Singh, "Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering," *Appl. Soft Comput.*, vol. 43, pp. 20–34, Jun. 2016.

[10] C. Bogner, B. T. Widemann, and H. Lange, "Characterising flow patterns in soils by feature extraction and multiple consensus clustering," *Ecol. Informat.*, vol. 15, pp. 44–52, May 2013.

[11] F. Saeed, N. Salim, and A. Abdo, "Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures," *Mol. Informat.*, vol. 32, no. 7, pp. 591–598, Jul. 2013.

[12] P. Blomstedt, R. Dutta, S. Seth, A. Brazma, and S. Kaski, "Modelling-based experiment retrieval: A case study with gene expression clustering," *Bioinformatics*, vol. 32, no. 9, pp. 1388–1394, May 2016.

[13] H. Rehioui and A. Idrissi, "New clustering algorithms for Twitter sentiment analysis," *IEEE Syst. J.*, vol. 14, no. 1, pp. 530–537, Mar. 2020.

[14] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy *k*-prototype clustering algorithm for mixed numeric and categorical data," *Knowl.-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2012, pp. 443–490.

[16] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," *Technometrics*, vol. 32, no. 2, pp. 227–229, 1988.

[17] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.

[18] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific–Asia Conf. Knowl. Discovery Data Mining*, 1997, pp. 21–34.

[19] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Boston, MA, USA: Kluwer, 1999.

[20] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved *k*-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013.

[21] S. P. Chatzis, "A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8684–8689, Jul. 2011.
[22] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul. 2002.
[23] C.-C. Hsu and Y.-C. Chen, "Mining of mixed data with application to catalog marketing," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 12–23, Jan. 2007.
[24] D. Lam, M. Wei, and D. Wunsch, "Clustering data of mixed categorical and numerical type with unsupervised feature learning," *IEEE Access*, vol. 3, pp. 1605–1613, 2015.
[25] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu, "Unsupervised evolutionary clustering algorithm for mixed type data," in *Proc. IEEE Congr. Evol. Comput.*, Barcelona, Spain, Jul. 2010, pp. 1–8.
[26] G. David and A. Averbuch, "SpectralCAT: Categorical spectral clustering of numerical and nominal data," *Pattern Recognit.*, vol. 45, no. 1, pp. 416–433, Jan. 2012.
[27] A. Foss, M. Markatou, B. Ray, and A. Heching, "A semiparametric method for clustering mixed data," *Mach. Learn.*, vol. 105, no. 3, pp. 419–458, Dec. 2016.
[28] J.-Y. Chen and H.-H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Inf. Sci.*, vol. 345, pp. 271–293, Jun. 2016.
[29] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," 2013, *arXiv:1304.5634*. [Online]. Available: http://arxiv.org/abs/1304.5634
[30] K. Kailing, H.-P. Kriegel, A. Pryakhin, and M. Schubert, "Clustering multi-represented objects with noise," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Berlin, Germany, 2004, pp. 394–403.
[31] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Mining*, Washington, DC, USA, Nov. 2004, pp. 1–8.
[32] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 129–136.
[33] A. Kumar, P. Rai, and H. Daumé, III, "Co-regularized multi-view spectral clustering," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[34] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, Jan. 2020.
[35] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, 2019, pp. 2563–2569.
[36] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Proc. Conf. 31st AAAI Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 2921–2927.
[37] F. Nie, J. Li, and X. Li, "Self-weighted multiview clustering with multiple graphs," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 2564–2570.
[38] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1774–1782, Jul. 2019.
[39] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971.
[40] Y. Yang, "An evaluation of statistical approaches to text categorization," *J. Inf. Retr.*, vol. 1, pp. 67–88, Apr. 1999.
[41] J. Ji, Y. Chen, G. Feng, X. Zhao, and F. He, "Clustering mixed numeric and categorical data with artificial bee colony strategy," *J. Intell. Fuzzy Syst.*, vol. 36, no. 2, pp. 1521–1530, Mar. 2019.

JINCHAO JI was born in Xuchang, Henan, China, in 1982. He received the bachelor's degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2007, and the Ph.D. degree in computer application technology from Jilin University, Changchun, China, in 2013.

From 2013 to 2015, he was a Postdoctoral Research Fellow of the Key Laboratory of Applied Statistics, MOE, Northeast Normal University, where he is currently a Lecturer. His research interests include machine learning, cluster analysis, and educational data mining.

RUONAN LI was born in Siping, Jilin, China, in 1996. She received the bachelor's degree in computer science and technology from the Changchun University of Technology, Changchun, China, in 2019. She is currently pursuing the master's degree with Northeast Normal University.

Her research interests include machine learning, cluster analysis, and educational data mining.

WEI PANG received the Ph.D. degree in computing science from the University of Aberdeen, in 2009. He is currently an Associate Professor with Heriot-Watt University, Edinburgh, U.K. He has authored more than 90 articles, including more than 30 journal articles. His research interests include bio-inspired computing, data mining, machine learning, and qualitative reasoning.

Dr. Pang was a recipient of the Best Paper Runner Up Award in the 12th International Conference on Advanced Data Mining and Applications (ADMA 2016) and the Best Paper Award in the 19th Annual UK Workshop on Computational Intelligence (UKCI 2019).

FEI HE was born in Nanning, Guangxi, China, in 1985. He received the bachelor's degree in computer science and technology and the Ph.D. degree in bioinformatics from Jilin University, Changchun, China, in 2007 and 2015, respectively.

He is currently a Lecturer with the School of Information Science and Technology, Northeast Normal University, China. His research interests include deep learning, bioinformatics, and biometrics.

GUOZHONG FENG was born in Kunshan, Jiangsu, China, in 1982. He received the bachelor's degree in mathematics and applied mathematics and the Ph.D. degree in probability theory and mathematical statistics from Northeast Normal University, Changchun, China, in 2004 and 2011, respectively.

From 2012 to 2015, he was a Postdoctoral Research Fellow with the Key Laboratory of Applied Statistics, MOE, Northeast Normal University, where he is currently an Associate Professor. He is also a Postdoctoral Scientist with the Department of Statistics, The George Washington University, Washington, DC, USA. His research interests include text mining and statistical machine learning.

XIAOWEI ZHAO was born in Shulan, Jilin, China, in 1984. She received the bachelor's and master's degrees in computer science and the Ph.D. degree in cell biology from Northeast Normal University, Changchun, China, in 2007, 2010, and 2013, respectively.

She is currently a Lecturer with the School of Information Science and Technology, Northeast Normal University. Her research interests include data mining, machine learning, and recommendation systems.