Leachable Component Clustering

Miao Cheng*
School of Computer Science
and Engineering
Guangxi Normal University
Guilin, Guangxi, China
Email: mcheng@mailbox.gxnu.edu.cn

Xinge You
School of Electronic Information
and Communication
Huazhong University of Science and Technology
Wuhan, Hubei, China
Email: youxg@hust.edu.cn

***Abstract*—Clustering attempts to partition data instances into several distinctive groups, while the similarities among data belonging to the common partition can be principally reserved. Furthermore, incomplete data frequently occurs in many real-world applications, and brings perverse influence on pattern analysis. As a consequence, the specific solutions to data imputation and handling are developed to conduct the missing values of data, and independent stage of knowledge exploitation is absorbed for information understanding. In this work, a novel approach to clustering of incomplete data, termed leachable component clustering, is proposed. Rather than existing methods, the proposed method handles data imputation with Bayes alignment, and collects the lost patterns in theory. Due to the simple numeric computation of equations, the proposed method can learn optimized partitions while the calculation efficiency is held. Experiments on several artificial incomplete data sets demonstrate that, the proposed method is able to present superior performance compared with other state-of-the-art algorithms.**

***Index Terms*—Incomplete data, leachable component clustering, Bayes alignment, calculation efficiency.**

I. Introduction

As an essential category of data analysis and handling methods, clustering aims to divide data instances into several separate data groups, by assuming the instances of each cluster share the common similarities of data patterns [1] [2]. As a consequence, different data points are to be organized regularly in accordance to certain partition rules, and normally, further handling can be achieved conveniently [3]. Until now, k-means has been the most popular solution to most clustering problems, due to its stable and outstanding performance [4], which is derived from the previous contribution of the classic Lloyd's method [5]. Furthermore, spectral clustering method [6] [7] has received broad attentions in recent years, which seeks for the approximate ideal solutions via decomposition of the spectral Laplacian. Nevertheless, it always suffers from the complexities of decomposition calculation, if large data is absorbed into clustering.

On the other hand, some existing works consider the data analysis with missing values [8] [9], which are in loss quite common during information convey and processing. The key-points of such problem, have been reduced to supplement or fix the broken data with certain repatching approaches in theory, and the resulting data can be adopted to further handling [10]. Since data blocks are normally represented as numeric matrices, it is referred as a matrix completion problem, and solved with optimization tools [11] [12]. Nevertheless, the shortcoming of high complexities prohibit it from calculation efficiency, while the concrete framework is hardly to be further developed. More frequently, incomplete information often occurs in big data handling [13] [14], e.g., database and recommendation systems, another category of solutions are referred to fill the incomplete information with data imputation methods. Furthermore, data imputation is adopted to computer vision to enhance component analysis with missing values [15] [16], and incremental learning of uncertain data has also been discussed [17]. As a consequence, some professional tools have been devised for data imputation, e.g., MICE [18], Amelia II [19]. In addition, multi-view clustering has attracted broad attentions in recent years [20] [21], which aims to learn the ideal partitions based on multiple representation of the common instances. It is noticeable that, the proposed method is also feasible for multi-view incomplete data analysis, owing to benefits of imputation models.

As a common limitation, clustering analysis of missing data are hardly to be addressed efficiently, as the two issues are independent problems and solved separately. Till now, there are quite a few solutions available for such issue, and most of them handle the clustering via a pre-step of data imputation [22] [23]. In order to improve the performance of data imputation approaches, there is a thirst for specific clustering solution to partitions of incomplete information [24] [25]. Based on fuzzy c-means clustering [26], Hathaway and Bezdek proposed a generalized clustering framework for incomplete data [27]. By devising a tolerant subspace clustering method, it is able to handle data with missing values flexibly [28]. In the literature, it has been reduced to be explained as optimization problems [29] [30] with specific objective functions that are defined accordingly [31]. In this work, a novel approach to clustering of incomplete data, termed leachable component clustering (LCC), is proposed. To distinguish from existing methods, the main contributions of this work are highlighted as below.

- The existing methods focus on representative clustering, while statistical patterns of incomplete data have been ignored. The proposed method seeks for optimized imputation, by exploiting the imputation models with respect to preservation of intrinsic distributions.
- The proposed imputation models aim to fulfill the lost el-

ements with the defined objectives, while fixed solutions can be obtained with calculation of the equations. Benefit from the efficiency of optimization, the ideal approximation can be calculated stably, while convergences can be achieved.
- The proposed framework can be considered as a unified solution to analysis of incomplete data, and leads to the possible extensions of further advances.

The rest of this paper is organized as follows. The background knowledge of self-expressive clustering method is introduced in section II, and the basic conception of latent component models are given in section III. Then, the main idea of the proposed LCC is given in section IV. The experimental results on several incomplete data sets are given in section V. Finally, the conclusion is draw in section VI.

II. Self-Expressive Clustering

Given a set of data instances $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d\times n}$, and the amount of clustering $c > 0$. The standard *k*-means clustering method attempts to assign each instance $x_i, \quad i = 1, 2, \cdots, n$ into c separate partitions, while the reconstruction errors of data can be minimized corresponding to the learned data means. Actually, it has been demonstrated that, it is equivalent for *k*-means and line reconstruction, and can be summarized as a common objective with a perspective of quadratic optimization problem [2]. Afterward, such idea is summarized as a self-expressive property [29] [30], while appended extensions can be developed conveniently. Without loss of generality, the self-expressive problem can be defined as

$$\begin{aligned} J_{SEC} &= \arg\min_{Z} \tfrac{1}{2}\|X - XZ\|^2 + \gamma \|Z\|_F^2 \\ s.t. &\quad Z^T 1 = 1, \quad 0 \le Z \le 1. \end{aligned} \tag{1}$$

Here, $Z \in \mathbb{R}^{n\times n}$ indicates the coefficient matrix that is able to approximately reconstruct the original data instances, γ denotes the balance parameter, while the sub-conditions can avoid the trivial solution and ensure the positive of Z. By extending the objective, the *k*-means clustering is to optimize the following function,

$$J_k = \arg\min_{Z} \frac{1}{2}\|X - M_x Z\|^2 + \gamma \|Z\|_F^2, \tag{2}$$

where $M_x \in \mathbb{R}^{d\times c}$ denotes the matrix consisting of data means of each cluster as one of its columns, and $Z \in \mathbb{R}^{c\times n}$ is the assigned indicator of partitions. Obviously, the *k*-means clustering aims to optimize the objective function by seeking for the ideal Z and updating M_x iteratively, which can approximate the original X.

III. Latent Component Models of Data Imputation

Data imputation aims to fullfil the missing values of data instances X, and certain mechnism is adopted to approximately learn the predicated values that can approach to the ground truth $\widetilde{X}$.

A. Principle Component Imputation Model

The well-known principal component analysis (PCA) [32] has been widely applied to learn the orthogonal components of the centered data block, and the main idea can be also explained as the ideal approximation of the original data in theory [33] [34] [35]. According to the self-expressive property, the orthogonal components of X and $\widetilde{X}$ nearly share the common bases. Furthermore, the learned orthogonal components are able to make ideal reconstruction of X, e.g.,

$$\begin{aligned} J_P &= \arg\min_{P} \left\|X - PP^T X\right\|^2 \\ s.t. &\quad P^T P = I, \end{aligned} \tag{3}$$

where $P \in \mathbb{R}^{d\times r}$ denotes the orthogonal bases of X. As a consequence, the reconstructed data can be adopted to learn the following bases accordingly, and the stable approximation is able to be obtained. With the reconstructed patterns $PP^T X$, it is believed that it holds the similar components with ground truth $\widetilde{X}$, and the ideal imputation of missing values can be estimated.

B. Self-Expressive Representation Learning

Though incomplete data brings difficulties in information analysis, it is still possible to be analyzed while the dilemma of handling of original data can be avoided. Without loss of generality, representation learning attempts to learn the representative patterns of the original data with respect to specific objectives [36] [37], e.g.,

$$f: \ x_i \to y_i, \quad i = 1, 2, \cdots, n. \tag{4}$$

As a consequence, the learned y_i are adopted to further handling, which are more optimistic for incomplete representation [38] [39].

More specifically, the similarity objectives of instances are normally referred [40], and thus, the objective of pseudo data can be defined as

$$J_S = \arg\min_{y} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|Sim\left(\widehat{x_{i,j}}, \widehat{x_{j,i}}\right) - Sim\left(y_i, y_j\right)\|^2. \tag{5}$$

Here, $Sim\left(\cdot,\cdot\right)$ denotes the similarity function between two instances. The $\widehat{x}_{i,j}$ and $\widehat{x}_{j,i}$ indicate the labeled instances corresponding to the valid features between two instances, which are respectively defined as

$$\widehat{x_{i,j}} = x_i \odot l_i \odot l_j \tag{6}$$

and

$$\widehat{x_{j,i}} = x_j \odot l_j \odot l_i, \tag{7}$$

where $\odot$ denotes the Hadamard product [35]. Furthermore, the l_i denotes the labels that represent the validation and missing values of instance x_i, e.g.,

$$\begin{aligned} l_{iq} &= \begin{cases} 1, & if\ x_{iq} \neq 0 \\ 0, & otherwise. \end{cases} \\ q &= 1, 2, \cdots, d. \end{aligned} \tag{8}$$

Here, x_{iq} denotes the q-th element of the incomplete vector x_i. As a consequence, the learned y_i can approximately preserve the original similarities of data pairs.

C. Information Volume Preservation Model

The main idea of information volume preservation (IVP) is based on the assumption that, the informative patterns of data actually hold a nearly constant volume corresponding to different partitions. Contrarily, IVP is able to contribute to pattern unfolding and distribution learning [41] [42]. In terms of this idea, it is to reserve the information contained in ground truth to be approximate to incomplete data, while reconstruction is adopted to estimate $\widetilde{x_i}$.

With respect to IVP model, the feature distances of incomplete data can be estimated and nearly approximate to the characteristics of ground truth, as the lost patterns can be weakened in the high-dimensional feature space. Particularly, the reconstruction of kernel components are adopted, and the information volumes $V_f\left(x_i\right)$ of kernels $h\left(\cdot,\cdot\right)$ associated with incomplete data $x_i,\quad i=1,2,\cdots,n$, can be defined as

$$\begin{aligned} &V_f\left(x_i\right)=\sum_{j=1}^{k}h\left(x_i\odot l_i\odot l_j, x_j\odot l_i\odot l_j\right),\\ &i=1,2,\cdots,n. \end{aligned} \tag{9}$$

Here, x_j denotes the k nearest neighbors of x_i. As a consequence, the dilemma of incomplete patterns can be alleviated, and IVP aims to preserve the obtained volumes in the fulfilled data that are to be estimated. Furthermore, it is necessary to make the data patterns be the explicit characteristics of normal instances, which can be further depicted with data similarities, e.g., euclidean or cosine distances. Accordingly, the objective of IVP model can be defined as

$$J_{IVP}=\arg\min_{\widetilde{x_i}}\left|V\left(\widetilde{x_i}\right)-V_f\left(x_i\right)\right|,\quad i=1,2,\cdots,n. \tag{10}$$

Thus, the ideal approximation can be learned, by solving the equations and repairing incomplete data alternately.

IV. Leachable Component Learning

Derived from information preservation model, a novel approach to data imputation is devised in this work. And the estimation of data distributions are exploited to predicate the lost patterns of incomplete data.

A. Bayes Alignment Model

Assume that the distribution parameters of fulfilled data are equivalent to the ones of ground truth, then it is feasible to learn the imputation by solving the latent equations of associated probabilistic models [43]. More specifically, the basic idea of the imputation model is based on the assumption that, the difference of data distributions hold equivalence between the incomplete and fulfilled instances, which can approximate to the distribution of ground truth. Distinguishingly, the Bayes alignment in the context, actually stands for the affiliation probabilities of each instance associated with certain data partitions, and are approximately represented by distributions. Nevertheless, it is worthwhile to highlight several calculation issues firstly. The sample means are to be calculated associated with the available values of incomplete data, and formally, the valid values of each instance is referred, which is defined as

$$\begin{aligned} &m_i=\tfrac{1}{\delta_i}\sum_{j=1}^{n}x_{ji}\cdot l_{ji},\quad i=1,2,\cdots,d.\\ &\delta_i=\sum_{j=1}^{n}l_{ji} \end{aligned} \tag{11}$$

Accordingly, the sample variance σ is to be calculated associated with the valid labels of data features. Then, the distribution of valid features of each data can be calculated according to normal distribution of samples. As a consequence, the distribution of samples with respect to a specific data x_i can be approximately calculated as

$$G_{m,\sigma^2}\left(x_i\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(x_i-m\odot l_i\right)^T\left(x_i-m\odot l_i\right)}{2\sigma^2}\right). \tag{12}$$

Note that, the obtained distribution $G_{m,\sigma^2}\left(x_i\right)$ is an estimated value of incomplete data associated with the available characteristics of data patterns.

As a consequence, it attempts to learn the ideal approximation $\widetilde{x_i}$ that can reserve the statistical distributions of each incomplete data, such as

$$J_{BA}=\arg\min_{\widetilde{x_i}}\left|G_{\widetilde{m},\widetilde{\sigma^2}}\left(\widetilde{x_i}\right)-G_{m,\sigma^2}\left(x_i\right)\right|,\quad i=1,2,\cdots,n. \tag{13}$$

Furthermore, the statistical parameters of fulfilled data can be calculated naturally, and obtain the approximate $\widetilde{m}$ and $\widetilde{\sigma^2}$. There are several available calculation solutions to obtain the latent values of incomplete data with respect to the objective of J_{BA}. Nevertheless, the objective of Bayes alignment imputation model can be solved with the simple calculation of equations.

B. Solution of Latent Leachable Learning

The leachable imputation models aim to learn the approximate incomplete data $x_i,\ i=1,2,\cdots,n$ with objectives of latent components. For the Bayes alignment model, it is to solve the following equation for each incomplete data,

$$\widetilde{x_i}\cdot\widetilde{x_i}-2\widetilde{m}\cdot\widetilde{x_i}+\widetilde{m}\cdot\widetilde{m}+2\widetilde{\sigma^2}\log\left(\sqrt{2\pi}\widetilde{\sigma}G_{m,\sigma^2}\left(x_i\right)\right)=0. \tag{14}$$

Note that, the objective is a standard quadratic equation, but the solution to the above equation normally exists in complex domain. Nevertheless, it can be approximated in real values with numeric rotations, and achievements are promised to be obtained with efficiency.

As a consequence, the ideal solution $\widetilde{x_i}$ to the proposed imputation model can be calculated as

$$\widetilde{x_i}=\widetilde{m}\pm\sqrt{2\widetilde{\sigma}^2\tau\left(\log\left(\sqrt{2\pi}\widetilde{\sigma}G_{m,\sigma^2}\left(x_i\right)\right)\right)}. \tag{15}$$

Here, $\tau\left(\cdot\right)$ denotes the absolute function, such as

$$\tau\left(x\right)=\begin{cases}\left|x\right|, & if\ x\neq 0\\ 0, & otherwise.\end{cases} \tag{16}$$

In addition, there obtained two alternative $\widetilde{x_i}$ for approximation in either model. During the first iteration, the $\widetilde{x_i}$ that is close to mean $\widetilde{m}$ is chosen as the solution of equation. Then, the ideal $\widetilde{x_i}$ is to be updated as the one that is close to the obtained $\widetilde{x_i}$ in the last iteration. Nevertheless, it has been shown that either $\widetilde{x_i}$ can be competent to learn the ideal results. After obtaining the approximate data, it is to update the parameters and repeatedly optimize the latent components during iterations.

Furthermore, the proposed method is able to recycling, due to its intrinsic relationship with data means. More specifically, the learned partitions can be the supposed affiliation of each instance as learning models in the next cycle, and the leachable components can be achieved with the fresh data groups. In other words, the parameters of distributions can be estimated and updated with the learned groups during each iteration, with respect to the instances that belong to the common partitions, as well as the lost patterns of instances. As a consequence, the fulfilled patterns can be achieved with repeated cycles, benefiting from the initially leachable learning.

C. Discussion

Compared with existing methods, the fixed solutions can be obtained with the proposed alignment model during each iteration, and the improved results are available in the further optimization steps. Nevertheless, the optimization convergence is hardly to be achieved, while the supplemented data can be approximated to the original data with respect to global distribution. As a consequence, the similarity function can be defined in accordance with Eq. (14), and the lost elements can be estimated by solving the approximate equation. Note that, the proposed supplementation models mainly rely on the solutions to line equations, which can be efficiently calculated with low computational complexities.

Similarly, the original IVP imputation model aims to solve the equation associated with the reservation of information volumes, e.g.,

$$\begin{aligned}
&\sum_{j=1}^{k}\left(\widetilde{x_i}\cdot\widetilde{x_i}-2\cdot x_j\cdot\widetilde{x_i}+x_j\cdot x_j\right)-\sum_{j=1}^{k}S_f\left(x_i,x_j\right)\\
&=k\cdot\widetilde{x_i}\cdot\widetilde{x_i}-2\cdot\widetilde{x_i}\cdot\sum_{j=1}^{k}x_j+\sum_{j=1}^{k}x_j\cdot x_j-\sum_{j=1}^{k}S_f\left(x_i,x_j\right)\\
&=0
\end{aligned} \tag{17}$$

where S_f denotes the similarities of incomplete data between x_i and x_j in feature space, which is defined as

$$S_f\left(x_i,x_j\right)=Sim_f\left(\widehat{x_i},\widehat{x_j}\right). \tag{18}$$

Nevertheless, it is hardly to achieve convergence in a common step, owing to independent optimization of each instance, which is necessary to calculate the information volumes associated with each instance. In practice, it is alleviated by setting a upper bound of iterations, and sampling is adopted to reduce the complexities.

TABLE I
THE DETAILS OF DATA SETS

Data	Partition	Dimensionality	Instance
Birds	20	50	3625
Firewall	3	11	3000
Flower	5	50	4323
Monkey Species	10	50	1098

V. EXPERIMENTS

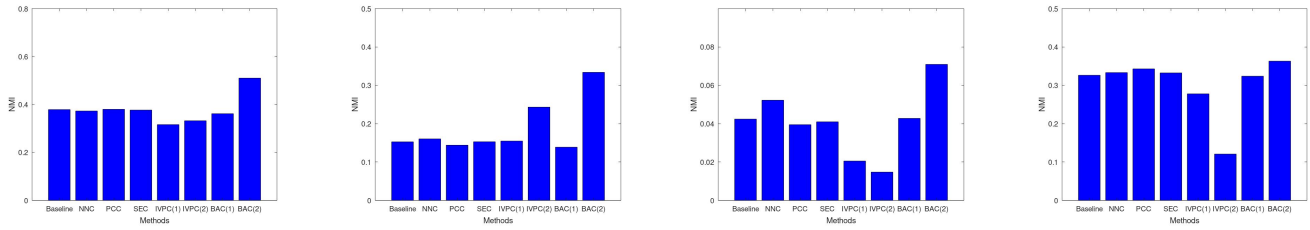
In this section, the performance of the proposed methods are to be evaluated on several artificial data sets. Four data sets are employed in the experiments, including Internet Firewall[1], 100 Birds[2], Flower Recognition[3], 10 Monkey Species[4]. All experiments are performed on the hardware of 2.9 GHz CPU with six cores and 16 GB RAM. During the experiments, partial instances of some data sets are employed, and the deep and reduced representations of image data sets are extracted, which consist of normalized patterns of 50 dimensions for each instance. For the Internet Firewall data set, the 1,000 instances of each category are randomly selected, and the image data of 20 categories in bird species are selected. In summary, the details of involved data sets are given in Tab. I.

Several state-of-the-art methods associated with data imputation of latent component models are involved to make a comparison of clustering performance, which are given as follows.

- Baseline *k*-means clustering [4]: The *k*-means clustering on raw incomplete data.
- Nearest neighborhood fulfilled clustering (NNC) [2]: The NN imputation based clustering of incomplete data. To reduce complexity, the average fulfillment from three random neighbors are preferred in the experiments.
- Principle components fulfilled clustering (PCC) [33] [34] [16]: The PC imputation based clustering of incomplete data.
- Self-expressive clustering (SEC) with incomplete data [29] [30]: The SEC method on raw incomplete data.
- Information volume preservation fulfilled clustering, namely IVPC(1) and IVPC(2): The proposed IVP imputation followed by the standard self-expressive and *k*-means clustering respectively.
- Bayes alignment fulfilled clustering, namely BAC(1) and BAC(2): The BA imputation followed by the self-expressive and *k*-means clustering respectively.

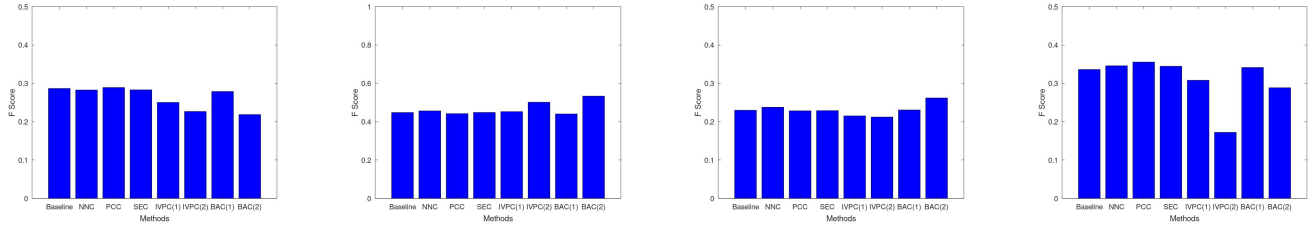
For each data set, *thirty* percent of whole elements are randomly selected and set to be null to make the incomplete data. Then, the data supplementation and standard SEC are performed by following each algorithm. Note that, the initial partitions of instances are quite important for optimized clustering, while random initialization is employed in the

[1]http://archive.ics.uci.edu/ml/datasets/Internet+Firewall+Data
[2]https://www.kaggle.com/gpiosenka/100-bird-species
[3]https://www.kaggle.com/alxmamaev/flowers-recognition
[4]https://www.kaggle.com/slothkong/10-monkey-species



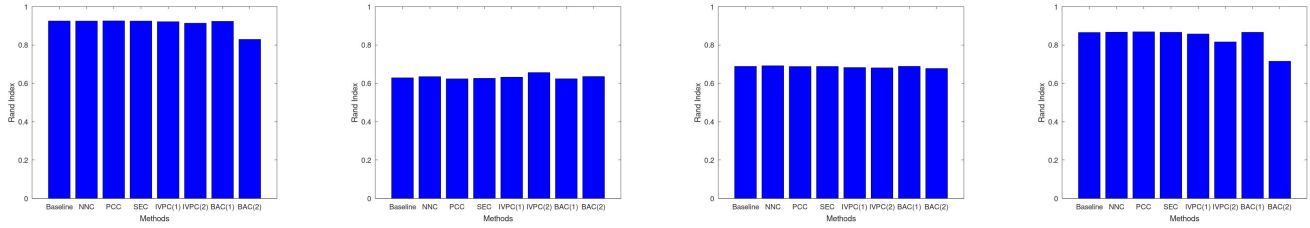
(a) (b) (c) (d)

Fig. 1. The obtained Normalized Mutual Information (NMI) of different algorithms on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.



(a) (b) (c) (d)

Fig. 2. The obtained F Score of different algorithms on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.



(a) (b) (c) (d)

Fig. 3. The obtained Rand Index of different algorithms on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.

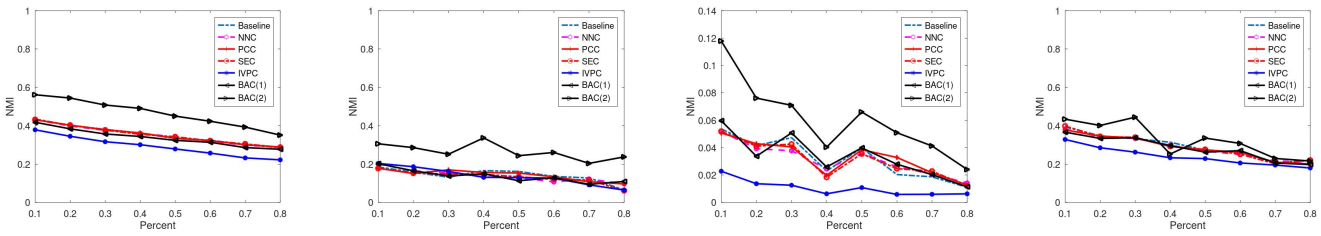
experiments. To alleviate the occasional sense, all experiments are repeated *five* times and the *average* results are calculated and recorded as the outputs. The obtained results of the first experiment are given in Fig. 1-3, corresponding to three clustering measures.

According to the experimental results, the best NMI results are obtained by BAC method on all four data sets, while better outputs are achieved with F measure. The obtained results of NNC and PCC are quite similar to each other, as well as the Baseline. In other words, these two methods produce the similar fulfillment for data imputation, and close performance are given for clustering. Furthermore, the outputs of IVPC are different from each other, corresponding to different clustering mechanism. And totally, k-means based IVPC is able to give better partitions compared with SEC in most cases, especially on the Firewall and Flower data sets. Though k-means based BAC is better with NMI and F measures on Firewall and Flower, the inferior results are obtained if the Rand Index measure is referred. Nevertheless, it is noticeable that, the obtained Rand Index of different methods are quite close to each other, and it is hardly to make conclusion.

In the second experiment, the performance of different methods with variational null elements are evaluated. More specifically, 0.1-0.8 percent of elements are randomly set to be null on four data sets, then all algorithms are repeatedly performed, and the average outputs are recorded as the results. In the experiment, the SEC based IVP method is performed to learn the data imputation associated with different percent of incomplete data. The NMI and F Score results are shown in Fig. 4-5, while Rand Index results are ignored due to the limited space.

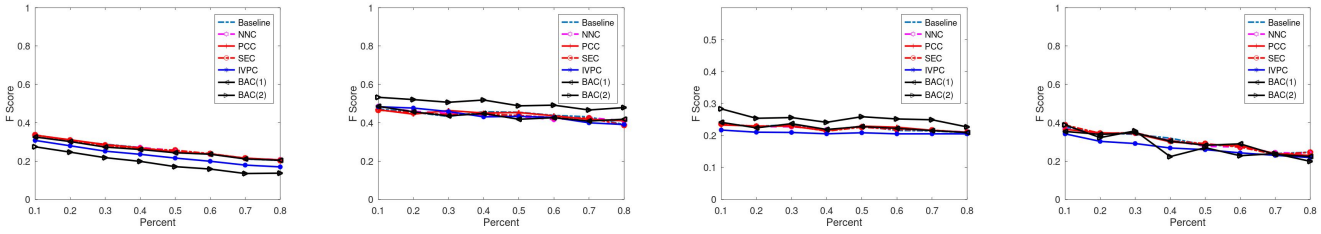
In terms of the results, the performance of most algorithms are quite similar to each other, and the fluctuating tendency of average results are reduced slowly as more unavailable elements are appended into experiments. Obviously, all methods are affected by the involved amount of null elements, and the results of NNC, PCC, and SEC are quite close, owing to the fact that the final outlets depend on either imputation and clustering, especially on Birds and Firewall. Compared with SEC based BAC, k-means BAC is superior than other methods, while IVPC is able to achieve comparable results.

In addition, the compuational costs associated with different percents of null elements are observed by recording the average time complexities of different methods, which is given



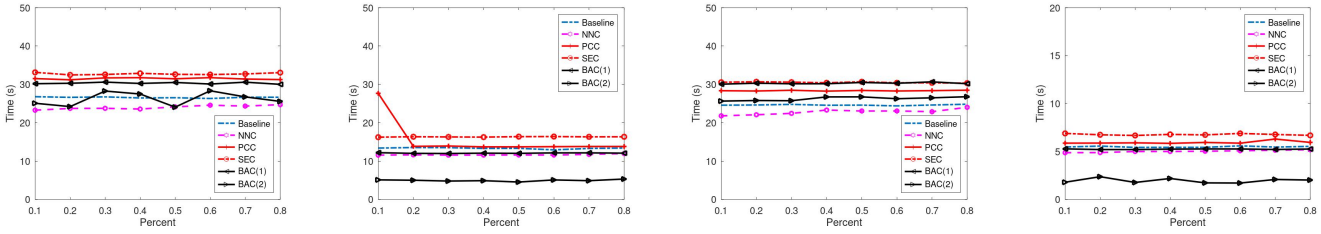
(a) (b) (c) (d)

Fig. 4. The obtained Normalized Mutual Information (NMI) of different algorithms associated with different percent of null elements on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.



(a) (b) (c) (d)

Fig. 5. The obtained F Score of different algorithms associated with different percent of null elements on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.



(a) (b) (c) (d)

Fig. 6. The time complexities (seconds) of different algorithms associated with different percent of null elements on four data sets. (a) Birds (b) Firewall (c) Flower (d) Monkey.

in Fig. 6. Nevertheless, IVPC is ignored due to its high complexities of calculation of information volumes for each instance. In terms of the observations, the proposed BAC can present the ideal performance if k-means is adopted to learn the partitions. Furthermore, stable time complexities are given by all methods as increasing parts of null elements, which discloses the efficiency of the involved methods. Since orthogonal decomposition is necessary for PCC, the most time complexities are required. Owing to random selection of neighbors, NNC method is able to present the optimistic efficiency on Birds and Flower data sets. Furthermore, SEC needs more computational times, as it is hardly to achieve stable convergence for self-expressive learning.

VI. Conclusion

Clustering attempts to partition data instances into several groups associated with the maximum similarities of common characteristics. Furthermore, incomplete data frequently occurs in many real-world applications, e.g, digital data conveying and information processing, and brings perverse influence on data handling with missing values. In this work, a novel approach to clustering of incomplete data is proposed, which normally fulfill the incomplete data by leachable component learning.

More specifically, the proposed model exploits the similarities of distributions between the repaired and incomplete patterns, and alignment framework is adopted to learn the fulfilled elements. Similar to predication of Bayes distribution, it is able to afford clustering of incomplete data with estimation of distribution parameters, and further extensions are possible with the proposed learning framework.

Experiments on diverse artificial data sets show that, the proposed method is able to give the outstanding performance compared with the state-of-the-art methods, while calculation efficiency is still held. The future works mainly focus on the advances of the proposed method to other topics of information handling with incomplete data, as well as the relative learning of pattern analysis. Besides, data fusion based leachable learning is quite valuable.

ACKNOWLEDGEMENT

The authors would like to thank anonymous reviewers for their constructive suggestions, and Firat University, General Dynamics, and Alexander Mamaev for providing data sets online. This work was partly supported by Innovation and Talent Foundation of Guangxi Province of China (RZ1900007485, AD19110154), and Natural Science Foundation of China (62172177). The corresponding author of this work is Miao Cheng.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. California, USA: Wiley, 2000.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[4] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[5] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[6] A. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2001.

[7] L. Z. Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2004, pp. 849–856.

[8] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Cambridge, UK: John Wiley & Sons, 2019.

[9] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Trans. System, Man, and Cybernetics*, vol. 9, no. 10, pp. 617–621, 1979.

[10] M. Cheng and A. C. Tsoi, "Crh: A simple benchmark approach to continuous hashing," in *Proc. Global Conference on Signal and Information Processing*, Orlando, USA, 2016.

[11] E. J. Candés and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[12] E. J. Candés and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[13] H. Nagashima and Y. Kato, "Method for selecting a data imputation model based on programming by example for data analysis," in *Proc. IEEE International Conference on Big Data*, Atlanta, USA, 2020.

[14] S. Song, Y. Sun, A. Zhang, L. Chen, and J. Wang, "Enriching data imputation under similarity rule constraints," *IEEE Trans. Knowledge and Data Engineering*, vol. 32, no. 2, pp. 275–287, 2018.

[15] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to sfm," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1051–1063, 2004.

[16] H. Y. Shum, K. Ikeuchi, and R. Reddy, "Principle component analysis with missing data and its application to polyhedral object modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 9, pp. 854–867, 1995.

[17] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Proc. European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[18] S. V. Buuren and K. G. Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[19] J. Honaker, G. King, and M. Blackwell, "Amelia ii: A program for missing data," *Journal of Statistical Software*, vol. 45, no. 7, pp. 1–47, 2011.

[20] W. Zhu, J. Lu, and J. Zhou, "Structured general and specific multi-view subspace clustering," *Pattern Recognition*, vol. 93, pp. 392–403, 2019.

[21] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proc. International Conference on Computer Vision and Pattern Recognition*, Nashville, USA, 2021.

[22] Q. Ma, Y. Gu, W. C. Lee, and G. Yu, "Order-sensitive imputation for clustered missing values," in *Proc. IEEE International Conference on Data Engineering*, Macao, China, 2019.

[23] Z. Charles, A. Jalali, and R. Willett, "Sparse subspace clustering with missing and corrupted data," in *Proc. IEEE Data Science Workshop*, Lausanne, Switzerland, 2018.

[24] E. R. Hruschka and N. F. F. E. E. R. Hruschka Jr, "Towards efficient imputation by nearest-neighbors: A clustering based approach," in *Proc. Austrilian Joint Conference on Artificial Intelligence*, Cairns, Australia, 2004.

[25] V. AUdigier, N. Niang, and M. R. Rigon, "Optimal clustering with missing values," in *Proc. International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Washington D. C., USA, 2018.

[26] M. Cheng, B. Fang, J. Wen, and Y. Y. Tang, "Marginal discriminant projections: An adaptable margin discriminant approach to feature reduction and extraction," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 1965–1974, 2010.

[27] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Trans. System, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 5, pp. 735–744, 2001.

[28] S. Gunnemann, E. Muller, S. Raubach, and T. Seidl, "Flexible fault tolerant subspace clustering for data with missing values," in *Proc. IEEE International Conference on Data Mining*, Vancouver, Canada, 2011.

[29] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE International Conference on Computer Vision*, Santiago, USA, 2015.

[30] Z. Kang, C. Peng, and Q. Cheng, "Twin learning for similarity and clustering: A unified kernel approach," in *Proc. AAAI Conference on Artificial Intelligence*, San Francisco, USA, 2017.

[31] V. AUdigier, N. Niang, and M. R. Rigon, "Generative adversarial nets," in *Proc. International Conference on Neural Information Processing*, Cambridge, USA, 2014.

[32] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[33] M. Cheng, Y. Y. Tang, and C. M. Pun, "Nonparametric feature extraction via direct maximum margin," in *Proc. International Conference on Machine Learning and Applications*, Honolulu, USA, 2011.

[34] M. Cheng, Z. Liu, H. Zou, and A. C. Tsoi, "A family of maximum margin criterion for adaptive learning," in *Proc. International Conference on Neural Information Processing*, Siem Reap, Cambodia, 2014.

[35] C. F. V. L. G. H. Golub, *Matrix Computations*, four edition ed. Baltimore, USA: John Hopkins University Press, 2013.

[36] N. Saeed, H. Nam, M. I. U. Haq, and D. B. M. Saqib, "A survey on multidimensional scaling," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–25, 2019.

[37] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Nature*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[38] D. L. Donoho, "Compressive sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[39] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[40] M. Cheng and X. You, "Adaptive matching of kernel means," in *Proc. International Conference on Pattern Recognition*, Milan, Italy, 2020.

[41] M. Cheng, "Ivp-ldl: Label distribution learning via preservation of information volumes," in *Proc. International Conference on Advanced Computational Intelligence*, Dali, China, 2020.

[42] M. Cheng, F. Zhou, H. Zhang, H. Zou, and J. Wu, "Function approximation for adaptive learning of label distributions," in *Proc. International Conference on Signal and Image Processing*, Nanjing, China, 2021.

[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Predication*, 2nd ed. Springer, 2011.