K-means for Datasets with Missing Attributes: Building Soft Constraints with Observed and Imputed Values

Diego P. P. Mesquita[1], João P. P. Gomes[1] and Leonardo R. Rodrigues[2] *

1- Federal University of Ceará - Computer Science Department
Rua Campus do Pici sn - Fortaleza-CE - Brazil

2- Institute of Aeronautics and Space - Electronics Division
Praça Marechal Eduardo Gomes, 50, São José dos Campos-SP - Brazil

Abstract. Clustering methods have a wide range of applications. However, the presence of missing attribute values on the dataset may limit the use of clustering methods. Developing clustering methods that can deal with missing data has been a topic of interest among researchers in recent years. This work presents a variant of the well known k-means algorithm that can handle missing data. The proposed algorithm uses one type of soft constraints for observed data and a second type for imputed data. Four public datasets were used in the experiments in order to compare the performance of the proposed model with a traditional k-means algorithm and an algorithm that uses soft constraints only for observed data. The results showed that the proposed method outperformed the benchmark methods for all datasets considered in the experiments.

1 Introduction

Clustering is an important machine learning problem that consists on partitioning a dataset into groups or clusters. Despite its wide range of applications, the usage of clustering methods may be limited by the presence of missing attributes on the dataset. The referred missing data problem is characterized by the absence of some features in several instances of the dataset. Values could be missing due to a number of reasons like sensor defects, data transmission problems and data storage problems.

To overcome this limitation, many authors proposed several clustering methods that can handle missing data. In [1] the authors identified two main alternatives to handle missing values: data imputation and marginalization.

In data imputation, missing values can be estimated based on a parametric/nonparametric estimation technique and a standard clustering method is applied. Examples of this approach can be found in [2], [3] and [4]. On the other hand, the marginalization approach discards the attributes that have missing values and apply standard clustering algorithms to the remaining features. Works based on this approach can be seen in [5], [6] and [7].

It is important to mention that both approaches present good results for different datasets and it is not possible to conclude that one has a superior

*This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq), grant 402000/2013-7.

performance in all cases. That said, an alternative method can be built by combining both approaches in a single algorithm.

In this work, we propose a variant of k-means that can handle missing values based on both marginalization and imputation frameworks. The proposed algorithm is based on the k-means with soft constraints (KSC) algorithm developed in [8]. In the KSC algorithm, an objective function is constructed based on the features that are available on all the dataset and soft constraints are built using the features that are missing in some feature vectors.

The proposed algorithm, referred as k-means with Soft Constraints on Observed and Imputed features (KSC-OI), adds an extra term to the objective function defined as a soft constraint for the imputed values. In the KSC-OI algorithm, the imputation is performed using the Incomplete Case k-Nearest Neighbors Imputation (ICkNNI) strategy [9], which is an imputation method based on the k nearest neighbors that use partial distances. Alternatively, it is possible to use any other imputation method.

The remaining Sections of this paper are organized as follows. Section 2 presents a brief description of the k-means clustering algorithm, as well as a description of the KSC algorithm proposed in [8]. Section 3 introduces the KSC-OI algorithm. Section 4 presents the results obtained in the experiments. Concluding remarks are presented in Section 5.

2 Theoretical Background

2.1 k-means Clustering

Given a number k of clusters, a similarity function $\psi(\cdot,\cdot)$ and a data set $\mathcal{X} = \{X_i\}_{i=1}^{N}$ comprising of n D-dimensional data points, the k-means algorithm aims to minimize the sum of the intra-cluster variances and can be summarized as follows:

- **Step 1:** Choose a set $\mathcal{C} = \{c_k\}_{k=1}^{K} \subseteq \mathcal{X}$ of initial centroids. An usual choice is to select k distinct members of the data set randomly.
- **Step 2:** Assign each data point X_i to the cluster $\mathcal{G}_m$ such that $m = \operatorname{argmin}_k \psi(X_i, c_k)$. The similarity is usually measured using the Euclidean distance.
- **Step 3:** Obtain the new set $\mathcal{C}$ of centroids by calculating the mean value of the members of each group.
- **Step 4:** Repeat steps 2 and 3 until convergence.

2.2 Clustering with Soft Constraints

Many classical clustering algorithms cannot deal with missing values. A common solution adopted in many applications is to fill in the missing values before using a clustering algorithm. One drawback of this approach is that the clustering algorithms will consider the imputed data as reliable as the observed data.

A variant of the k-means algorithm that uses soft constraints in order to avoid imputation has been proposed in [8]. In this algorithm, referred as KSC, the set of features $F = \{1, \ldots, D\}$ is split between subsets F_o and F_m, such that $p \in F_o$ if and only if $X_{i,p}$ is an observed value for all data points. Then, for each pair (X_i, X_j) of data points in $\mathcal{X}$, is created a soft constraint of strength of $s_{i,j}$, given by

$$s_{i,j} = \begin{cases} -\sqrt{\sum_{f \in F_m} (X_{i,f} - X_{j,f})^2} & \text{if } M_i = M_j = \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $M_i, M_j \subseteq F$ are respectively the indexes of the missing features in X_i and X_j. Note that soft constraints are defined only between complete data items, for data items with missing values no constraints will be created. A constraint between X_i and X_j is said to be violated if both data points reside in the same cluster.

The main idea behind the KSC algorithm is to cluster the items based on the features in F_o with soft constraints based on the features in F_m. In order to do so, it tries to minimize

$$g = w_1 \sum_{k=1}^{K} \sum_{X_i \in \mathcal{G}_k} \|X_i - c_k\|_2^2 + w_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{i,j} s_{i,j}^2 \tag{2}$$

where w_1 and w_2 are user defined hyper-parameters and $\delta_{i,j}$ is a binary variable that assumes value 1 if X_i and X_j are assigned to the same cluster and 0 otherwise.

The KSC algorithm is described as follows:

- **Step 1:** Choose a set $\mathcal{C} = \{c_j\}_{j=1}^{k} \subseteq X$ of initial centroids.
- **Step 2:** Assign each data point X_i to cluster $\mathcal{G}_m$ such that

$$m = \underset{1 \le k \le K}{\operatorname{argmin}}\, w_1 \|X_i - c_k\|_2^2 + w_2 \sum_{j=1}^{N} \delta_{i,j} s_{i,j}^2$$

- **Step 3:** Obtain the new set $\mathcal{C}$ of centroids by averaging over the features in F_o for each cluster.
- **Step 4:** Repeat steps 2 and 3 until convergence.

3 The KSC-OI Algorithm

Although KSC takes into account information that would be lost if a common clustering algorithm was used in the F_o space, it ignores information from partially complete data points. It is also important to remember that KSC completely rules off the possibility of data imputation.

In an effort to solve these problem, we present the KSC-OI algorithm. The proposed method extends the KSC algorithm by adding new soft constraints based on already present entries that were ignored and also on imputed values.

For each X_i, let $\tilde{X}_i$ be its imputed version. Along with the original KSC soft constraints, for each pair (X_i, X_j) of data points in X, we introduce a constraint of strength $\tilde{s}_{i,j}$, given as follows:

$$\tilde{s}_{i,j} = \begin{cases} -\sqrt{\sum_{f \in F_m} (\tilde{X}_{i,f} - \tilde{X}_{j,f})^2} & \text{if } M_i \cup M_j \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The objective function $\tilde{g}$ for KSC-OI is then given by

$$\tilde{g} = w_1 \sum_{k=1}^{K} \sum_{X_i \in \mathcal{G}_k} \|X_i - c_k\|_2^2 + w_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{i,j} s_{i,j}^2 + w_3 \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{i,j} \tilde{s}_{i,j}^2 \tag{4}$$

and the algorithm to do so is summarized as follows:

- **Step 1:** Choose a set $\mathcal{C} = \{c_j\}_{j=1}^{k} \subseteq \mathcal{X}$ of initial centroids.
- **Step 2:** Assign each data point X_i to cluster $\mathcal{G}_m$ such that

$$m = \underset{1 \leq k \leq K}{\operatorname{argmin}}\, w_1 \|X_i - c_k\|_2^2 + w_2 \sum_{j=1}^{N} \delta_{i,j} s_{i,j}^2 + w_3 \sum_{j=1}^{N} \delta_{i,j} \tilde{s}_{i,j}^2$$

- **Step 3:** Obtain the new set $\mathcal{C}$ of centroids by averaging over the features in F_o for each cluster.
- **Step 4:** Repeat steps 2 and 3 until convergence.

4 Experiments

In order to asses the performance of the KSC-OI algorithm, experiments with four real world datasets were conducted. Table 1 presents the characteristics of the datasets that were used.

Table 1: Data sets description

Dataset	**Size**	**Features**	**# clusters**
Glass	214	10	2
Wine	178	13	3
Iris	150	4	3
Breast Cancer	569	30	2

On the experiments, the performance of the proposed algorithm was compared to the k-means and KSC algorithms. The rand index was chosen as the performance metric. Weights w_1 and w_2 on the KSC algorithm were set to 0.5. Weights w_1, w_2 and w_3 on the KSC-OI algorithm were set to 1/3. Table 2 presents the rand index statistics for 10 executions of each algorithm in each dataset.

Table 2: Performance comparison in different data sets

	***k*-means**	**KSC**	**KSC-OI**
Glass	0.246 ± 0.089	0.590 ± 0.062	0.690 ± 0.012
Wine	0.437 ± 0.022	0.477 ± 0.004	0.497 ± 0.004
Iris	0.358 ± 0.017	0.425 ± 0.001	0.438 ± 0.011
Breast Cancer	0.498 ± 0.024	0.568 ± 0.013	0.630 ± 0.032

As can be noticed, the KSC algorithm outperformed k-means for all datasets. This fact shows the impact of using soft constraints on observed data. It is also possible to notice that the proposed algorithm achieved better results than the KSC in all datasets. It is more noticeable for the Glass and the Breast Cancer datasets.

It is worth pointing that KSC-OI can be seen as a generalization of both k-means and KSC algorithms. Setting $w_3 = 0$ leads the objective function of KSC-OI to be equivalent to the objective function of KSC. Setting both $w_2 = w_3 = 0$ leads the objective function of KSC-OI to be equivalent to the objective function of k-means.

5 Conclusions

A variant of the k-means clustering algorithm for datasets with missing data values was introduced. The proposed algorithm is based on a soft constraints framework and builds soft constraints for observed and imputed data.

The results obtained in the experiments using four real world public databases showed that the proposed algorithm outperformed the traditional k-means algorithm and the KSC algorithm. This can be explained by the fact that KSC-OI algorithm combines the imputation and marginalization approaches, which are the two main strategies to deal with missing values found in literature. It is worth noticing that the parameters of the KSC-OI algorithm can be adjusted so that the final objective function is more sensible to the constraints related to the imputed or the observed features.

Future works may include strategies to set w_1, w_2 and w_3 for a given clustering problem. Additionally, we intend to formulate extensions of other clustering algorithms to the same missing data framework.

Acknowledgments

The authors acknowledge the support of CNPq (Grant 402000/2013-7).

References

[1] Phil Green, Jon Barker, Martin Cooke, and Ljubomir Josifovski. Handling missing and unreliable information in speech recognition. In *In Proc*, 2001.

[2] Alan Wee-Chung Liew, Ngai-Fong Law, and Hong Yan. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12:498–513, 2010.

[3] José M. Jerez, Ignacio Molina, Pedro J. GarcÃa-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50:105–115, 2010.

[4] P. Saravanan and P. Sailakshmi. Missing value imputation using fuzzy possibilistic C means optimized with support vector regression and genetic algorithm. *Journal of Theoretical and Applied Information Technology*, 72:34–39, 2015.

[5] Giulio Genovese, Robert E. Handsaker, Heng Li, Eimear E. Kenny, and Steven A. McCarroll. Mapping the human reference genome's missing sequence by three-way admixture in latino genomes. *American Journal of Human Genetics*, 93:411–421, 2013.

[6] Soren Fons Vind Nielsen and Morten Morup. Non-negative tensor factorization with missing data for the modeling of gene expressions in the human brain. In *Proceedings of the 2014 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2014.

[7] Sjors H.W. Scheres, Roberto Melero, Mikel Valle, and Jose-Maria Carazo. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure*, 17:1563–1572, 2009.

[8] K. Wagstaff. Clustering with missing values: No imputation required. In *Proceedings of the Meeting of the International Federation of Classification Societies*, pages 649–658, 2004.

[9] Jason Van Hulse and Taghi M. Khoshgoftaar. Incomplete-case nearest neighbor imputation in software measurement data. *Information Sciences*, 259:596–610, 2014.