Missing values in dissimilarity-based classification of multi-way data

Diana Porro-Muñoz[2], Robert P. W. Duin[2] and Isneri Talavera[1]

Advanced Technologies Application Center (CENATAV), Cuba,
and Pattern Recognition Lab, TU Delft, The Netherlands,
dporro@gmail.com, r.duin@ieee.org, italavera@cenatav.co.cu

Abstract. Missing values can occur frequently in many real world situations. Such is the case of multi-way data applications, where objects are usually represented by arrays of 2 or more dimensions e.g. biomedical signals that can be represented as time-frequency matrices. This lack of attributes tends to influence the analysis of the data. In classification tasks for example, the performance of classifiers is usually deteriorated. Therefore, it is necessary to address this problem before classifiers are built. Although the absence of values is common in these types of data sets, there are just a few studies to tackle this problem for classification purposes. In this paper, we study two approaches to overcome the missing values problem in dissimilarity-based classification of multi-way data. Namely, imputation by factorization, and a modification of the previously proposed Continuous Multi-way Shape measure for comparing multi-way objects.

Key words: missing values, multi-way data, dissimilarity representation

1 Introduction

Classification problems are very common in most research areas, and a suitable representation of objects plays an important role in this task. However, even when this representation is found, problems like the absence of values for some of the measured features can affect the accuracy of classifiers. There can be several reasons for data to be missing. Namely, equipments malfunctioning, data were not entered correctly or data just do not exist for some objects, etc. In other cases, missing values are not actually present in the obtained data. Nonetheless, they are inserted as a postprocessing in order to make the data more appropriate to be described for some specific models [1, 2].

For many applications e.g. neuroinformatics, chemometrics, data sets can have a multi-dimensional structure e.g. $objects \times frequencies \times time$, instead of the simple vector representation. These structures are often richer in information, thus advantageous for many purposes as classification. Therefore, it is important to employ proper tools in order to analyze them. As in the two-dimensional case, these types of data may be affected by the presence of missing values. For multi-way data, different behaviors for missing data can be observed [3, 4] (See Fig. 1). The simplest case is when missing values are random without any pattern, denoted as RMV in [3]. Another common pattern is when complete fibers i.e. rows or tubes (See Fig. 1) are missing at random (RMF). A third pattern is when missing values are systematic for all objects (SMV) i.e. the same values are missing for all objects. In contrast with the two-way case, there

is just a limited research addressing the problem of missing values in multi-way data. Most of the related studies are dedicated to the robustness of factorization methods. Examples of the most common methods are PARAFAC algorithms based on Expectation Maximization - Alternating Least Squares [3] and based on the Levenberg - Marquadt method known as INDAFAC [3]. A more recent development is the CP-WOPT algorithm [5]. Other extensions of the multi-way methods, like TUCKER3, for dealing with missing values can be found in [6, 7, 4]. However, these methods are based on seeking accuracy in the obtained factor.

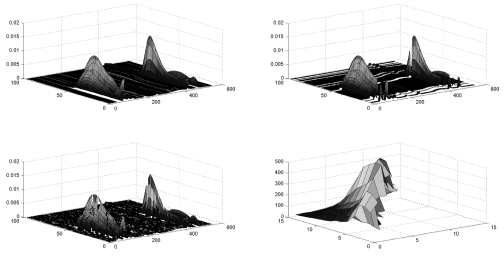

Fig. 1: Different patterns of missing values in a 2D object: RMF (top-left) Rows and (top-right) Tubes, (bottom-left) RMV and (bottom-right) SMV with an example of excitation-emission fluorescence data.

In this paper, we make a study on how to deal with missing values with the aim of minimizing the error function in the classification of multi-way data. We will use the Dissimilarity Representation (DR) [8] approach recently extended for the classification of multi-way data [9]. Roughly speaking, in this approach, (dis) similarities between objects are used as new features to describe them in a new space. Classifiers can be used in this space as in the traditional feature space. One of the approaches to deal with missing data in this case could be to reconstruct the data by a factorization method before the computation of the dissimilarity matrix. Another variant for dealing with missing data, particularly for the DR approach, consists in modifying the dissimilarity measure. With this purpose, we introduce a modification of the dissimilarity measure that will be used here such that it can deal with missing values.

The paper is organized as follows. The DR approach is briefly explained in Section 2. A description and comparative analysis of the studied approaches is presented in Section 3. Section 4 is dedicated to the experiments and discussion. Conclusions are presented in Section 5.

2 Dissimilarity Representation

The Dissimilarity Representation (DR) [8] approach has been introduced for classification purposes. It consists in a representation of objects by their (dis) similarities to a set of prototypes of each class identified in the problem at hand. One of the advantages of this approach is that it can be obtained from any representation of objects e.g. graphs, multi-dimensional objects, as long as a suitable

measure is used. Moreover, this approach allows introducing discriminative context information that helps for a better discrimination of objects.

Let us define the Dissimilarity Space (DS) approach, given a t-way array $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_t}$ where each object is represented by a (t-1)-dimensional array, a representation set $\underline{R}(\underline{R_1}, ..., \underline{R_h})$ where h is the number of prototypes, and a dissimilarity measure d [8, 9]. A mapping $\phi(\cdot, \underline{R}) : \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}} \rightarrow \mathbb{R}^h$ is done, such that every object $\phi(\underline{Y_i}, \underline{R}) = [d(\underline{Y_i}, \underline{R_1}), d(\underline{Y_i}, \underline{R_2}), ..., d(\underline{Y_i}, \underline{R_h})]$ is associated by its dissimilarities to all objects in $\underline{R}$. Hence, a dissimilarity matrix $\mathbf{D}(\underline{Y}, \underline{R})$ is obtained, which is used to build a classifier in the correspondent dissimilarity space of dimension h. The prototypes are usually the most representative objects of each class, $\underline{R} \subseteq \underline{Y}$ or $\underline{Y}$ itself. Any traditional classifier can be built in the dissimilarity space as in the feature space. Few work has been done to treat missing data in the DR approach. In [10], two alternatives for dealing with missing values in the dissimilarity representation-based classification are proposed. However, this work is only based on 2D data, where objects are represented by vectors in the feature space. It does not fit multi-way data. In this paper, we study two alternatives for classifying incomplete multi-way data by using the DR. The first approach is based on completing the multi-way data with a factorization method before computing the dissimilarity matrix (See Section 3). The second alternative consists in adapting the dissimilarity measure, such that dissimilarities between objects are obtained from the available information only.

The data sets to be studied here have a continuous nature. The characteristic shape of the surfaces for each class of objects is an important discriminative property of these type of data. Moreover, the information from the multi-way structure should be taken into account. Recently, the Continuous Multi-way Shape (CMS) [11] was introduced with this purpose. It consists in the comparison of multi-way objects based on the differences of their multi-way shape, considering the connectivity that exists between the neighboring points in the different directions. Differences between the gradients of the surfaces of these objects are computed, based on the application of linear filters by convolution. Thus, given $\underline{Y_a}$, $\underline{Y_b}$ two multi-way objects from a multi-way data set $\underline{Y}$, the dissimilarity measure CMS can be defined as:

$$d_G(\underline{Y_a}, \underline{Y_b}) = \left\| \sum_{i=1}^{f} \underline{Y_a} * \underline{G_\sigma} * \underline{H_i} - \underline{Y_b} * \underline{G_\sigma} * \underline{H_i} \right\|_F \quad (1)$$

where $\| \cdot \|_F$ is the Frobenius norm for tensors [12], $*$ is the convolution operator [13], $\underline{G_\sigma} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}}$ a Gaussian convolution kernel to smooth the data, $\underline{H_i} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}}$ is a partial derivative kernel and f is the number of partial derivatives in the different directions to obtain the gradient. The modification of the CMS measure for missing values will be given in the next Section.

3 Dealing with missing values in multi-way data

3.1 Factorization-based estimation

Factorization methods are very common for the analysis of multi-way data. They are used to extract and model their underlying structure. These methods are affected by the missing values, as data can be improperly analyzed. Therefore, creating robust methods to missing data has been one of the main tasks in the

development of factorization methods [4]. Such is the case of the PARAFAC [3, 4], which is one of the most used methods for multi-way data analysis.

Given the three-way array $\underline{\mathrm{Y}}$ of dimensions $I \times J \times K$, the PARAFAC model (decomposition) [14, 7] can be expressed by the factor matrices $A(I \times F)$, $B(J \times F)$ and $C(K \times F)$, such that $\underline{y}_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}$, where $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$ and F is the number of selected factors. In principle, factorization methods handle missing data with the aim of obtaining the most accurate data model. However, once the factorization has been computed, the resulting factor matrices can be used to reconstruct the original data and missing values are then estimated. Thus, a multi-way without missing values $\underline{\hat{Y}} \approx \underline{Y}$ can be obtained, using the information from the whole multi-way structure.

PARAFAC-Alternating Least Squares (ALS) [3] and CANDECOMP/PARAFAC Weighted OPTimization (CP-WOPT) [5] are two of the main algorithms for fitting the PARAFAC model with missing data. PARAFAC-ALS works well for small amounts of missing data and it is very simple and fast. However, it may suffer of slow/no convergence as the amount of missing values increases [3]. It also depends on the patterns of the missing values. CP-WOPT is a scalable algorithm, which is based on direct non-linear optimization to solve the least squares problem. This algorithm has shown to work well even with 70% of missing data and it is fast. However, this method has not been studied for missing data with a SMV pattern. Both algorithms will be used here as means of estimation of missing values for the classification of incomplete multi-way data sets.

3.2 Ignoring missing values in DR: adjustment of CMS measure

An alternative for dealing with missing values in the DR approach is to compute proximities on available data only. However, this approach depends on the measure to be used i.e. the definition of each measure has to be adapted for this purpose, which is not always straightforward. In this paper, the adaptation of the CMS measure will be explained. Although the CMS measure was proposed for multi-way data in general, in this paper we will focus on three-way data only.

In this measure, missing values will be treated in the first step i.e. Gaussian filter. The idea is to use a filter that will only process the non-missing values in the analyzed window. In practice, if we have a matrix $\mathbf{Y}$ and a 2D filter kernel $\mathbf{G}$, the result of applying the filter $\mathbf{G}$ (or any other filter) at each position of matrix $\mathbf{Y}$ is defined as:

$$\mathbf{Y}'(u,v) = \sum_{k=-P}^{P} \sum_{l=-P}^{P} \mathbf{Y}(u-k, v-l) \cdot \mathbf{G}(k,l)$$

where $2P+1$ is the size of the filter in both the horizontal and vertical directions of the convolution kernel $\mathbf{G}$. So, suppose we are analyzing a part of the data with q missing values. The filter is only applied to the $(2P+1)^2 - q$ non-missing values. In such case, as the number of summed values are less, the filtering result S for the analyzed position will not correspond to that if the data was complete. Therefore, a normalization like $S' = \frac{S}{(2P+1)^2-q} \cdot (2P+1)^2$ should be applied. If S' is used as the filtering result, instead of S, we are doing an implicit estimation of the missing values. That is, we are assuming that each missing value contributes in the filtering result S' with a value of $\frac{S}{(2P+1)^2-q}$. However, this can be considered a drawback, since the implicit estimation of the

missing value can change according to the position of the filter on the 2D matrix. When the amount of missing values is large, it could happen that all values in the window of the analyzed point are missing. In such situation, the previous adaptation does not work, it assigns NaN to the analyzed point. In this case, the idea is then to omit these points when objects are compared.

4 Experimental setup and discussion

The main goal of the following experiments is to evaluate how the factorization methods and the proposed adaptation of the CMS measure contribute to the DR-based classification of incomplete multi-way data. With this purpose, 4 different three-way continuous data sets are used. The first data set is private and it comes from 1200 patches of 1024×1024 pixels of 36 colon tissue slides from Atrium hospital in Heerlen, The Netherlands. Patches were filtered with Laplace filters in 90 different scales using $\sigma = 2.\hat{}[0.1 : 0.1 : 9]$. The log-squares of the results are summarized in 60 bin normalized histograms with bin centers $[-50 : 1 : 9]$. Thus, a 90×60 array is obtained for every patch, leading to a three-way array of $1200 \times 90 \times 60$. The patches are labeled in two classes: Normal and Tumor. The second data set consists in metabolite data containing HPLC measurements of commercial extract of St. John's wort [15]. HPLC-PDA (HPLC with photo-diodo array detection) profiles were obtained from 24 different examples of St. Johns' wort from several continents: Africa (8 objects), Asia (6 objects), Europe (45 objects) and North America (30 objects). The final three-way data has a size of $89 \times 97 \times 549$. The third and fourth data sets are from public domains and they are both obtained by Fluorescence spectroscopy. Fluorescence excitation-emission measurements are used because they are known to reflect important properties of the fermentation process. The first data set consists of a training set of size $323 \times 15 \times 15$ and a test set of $53 \times 15 \times 15$. The two variable modes correspond to excitation and emission wavelengths respectively. The classification problem consists in determining the quality (Low or High) of a process according to the enzyme activity [1]. The other data set has a size of $67 \times 11 \times 13$ and the purpose is to determine the age range of a Parma ham sample: raw (0 months), salted (3 months), matured (11 and 12 months) and aged (15 and 18 months) [2].

In the first two data sets there are no missing values, but these were generated artificially to test the methods. For each of them, 10 new data sets were first created by inserting various amounts $(1 - 5, 10, 20, 30, 50, 70\%)$ of missing values in the whole data set. Thus, all objects have the same probability of having missing values and the amount per object is completely random. This procedure was done for RMV and FMV patterns. Hence, we have generated in total 30 new data sets from the original ones. To avoid that results are influenced by a specific random pattern, we repeated the previous configurations 5 times for each of the two data sets. In the case of autofluorescence data, missing values were originally introduced as part of a preprocessing. Before analysis, values for emission wavelengths below excitation wavelengths should be deleted. Therefore, there is a systematic pattern of missing values for all objects, corresponding to those of the specified excitation/emission wavelengths.

The imputation by factorization and the modified CMS measure explained in Section 3 are evaluated on all data sets. There are different methods for the selection of the number of components in the factorization-based methods [7]. However, as in our case the interest is to reconstruct the original data as good as possible, we will use the residuals evaluation criteria. This consists on trying to find a minimum sum of squares of errors in the approximation of the

non-missing values. In all cases, classifiers performed better for those models that fulfilled the previous criteria. Results are given in terms of classification error. The Regularized linear discriminant classifier was used on the dissimilarity space [8]. To find the regularization parameters of RLDC, an automatic regularization (optimization over training set by cross-validation) process was done. For the different data sets, experiments were carried out differently. For small data sets (Parma ham and St John's), classification errors were obtained in a 10 times k-fold cross-validation (CV). The Enzyme data comes with training and test set. In the case of Colon data, 10 different training (90%) and test (10%) sets were randomly chosen and the error values were averaged. Experiments for the 5 repetitions of each of the configurations were averaged.

In the DR approach, for the small data sets, the representation set has the same size of the training set obtained in each fold of the cross-validation procedure. For the Colon data set, a representation set of 550 prototypes was randomly chosen for each generated training set. In the case of the Enzyme data set, 100 prototypes were also randomly chosen from the training set.

Table 1: Classification errors of Colon Cancer and St John's data sets after treatment of missing values. Results for different percents and patterns of missing data are shown. The baseline errors with the complete data are 0.095 and 0.02 respectively.

Colon Cancer data set										
	Random missing values (%)									
Methods	1	2	3	4	5	10	20	30	50	70
PARAFAC	0.27	0.3	0.32	0.32	0.32	0.33	0.34	0.32	0.36	0.39
CP-WOPT	0.18	0.2	0.22	0.22	0.26	0.26	0.28	0.28	0.36	0.36
Adapted CMS	0.12	0.13	0.14	0.14	0.14	0.17	0.19	0.21	0.28	0.30
	Complete tubes missing(%)									
PARAFAC	0.30	0.30	0.31	0.31	0.31	0.31	0.31	0.32	0.33	0.40
CP-WOPT	0.2	0.2	0.2	0.26	0.22	0.24	0.24	0.26	0.28	0.36
Adapted CMS	0.14	0.14	0.14	0.14	0.13	0.16	0.17	0.19	0.24	0.29
	Complete rows missing(%)									
PARAFAC	0.31	0.31	0.31	0.32	0.32	0.32	0.32	0.34	0.38	0.4
CP-WOPT	0.19	0.24	0.2	0.22	0.24	0.22	0.22	0.28	0.28	0.34
Adapted CMS	0.14	0.15	0.15	0.15	0.15	0.14	0.15	0.18	0.19	0.25
St John's data set										
	Random missing values (%)									
Methods	1	2	3	4	5	10	20	30	50	70
PARAFAC	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CP-WOPT	0.03	0.03	0.03	0.03	0.04	0.04	0.10	0.10	0.10	0.12
Adapted CMS	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.07	0.17	0.26
	Complete tubes missing (%)									
PARAFAC	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.06
CP-WOPT	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05
Adapted CMS	0.03	0.03	0.03	0.04	0.03	0.03	0.04	0.08	0.24	0.41
	Complete rows missing (%)									
PARAFAC	0.05	0.05	0.05	0.04	0.05	0.06	0.07	0.11	0.13	0.23
CP-WOPT	0.04	0.04	0.04	0.04	0.05	0.07	0.07	0.09	0.12	0.19
Adapted CMS	0.02	0.03	0.03	0.03	0.03	0.06	0.13	0.22	0.32	0.46

Table 1 summarizes the classification errors of two of the data sets after reconstructing the data. The classification errors on the complete data sets are used as a baseline for the comparison. Factorization-based methods work well in general when they converge, like in the case of St John's. However, this is not the case when convergence is not reached. It can be observed in Colon data set that performance of the classifier is bad for all patterns of missing values. It is actually the worst result. In this case, both algorithms took long to converge

and for large amounts of missing data convergence was never reached. There is a slightly improvement of CP-WOPT based results over those of PARAFAC-ALS, specially for large amounts of missing data, as expected. It has to be noticed that for large amounts of missing values these methods are stable. Nonetheless, even when the stability of these methods (when they converge) for different amounts of missing data is very attractive, their slow/no convergence problem is a strong drawback when comparing methods to reconstruct the missing data.

Let us analyze the adapted CMS measure directly applied on the incomplete data. It has to be remarked that for small amounts of missing data $(1 - 5\%)$, classifiers performance is comparable with the baseline classification error. In fact, for St John's data set, the baseline classification error is reached. A very attractive characteristic of the modified measure is that without a preprocessing step i.e. imputation, approximation, it has shown to work well with small amounts of missing data. Therefore, it can be a good option for these types of problems. Good performances can be obtained without the need of the extra computational cost of the imputation process. However, when the amount of missing values is large (usually above 10%), the classifier seems to loose stability and the performance gets drastically worse the more missing values are added. This could be explained by the fact that when there are many contiguous windows missing, the Gaussian filter cannot deal with it, too much information is lost and the idea of derivatives is kind of pointless. In these cases, the use of an imputation method e.g. interpolation is recommended. In general, for the two analyzed patterns of missing values, that is RMV and FMV, all methods behaved similar. It can be observed that the type of pattern of the missing points did not have a strong influence in the performance of the methods. The main disturbing factor was the amount of missing values.

Table 2: Classification errors of Enzyme and Parma ham data set after treatment of systematic missing values.

Methods	Data sets	
	Enzyme	Parma ham
PARAFAC	0.09	0.08
CP-WOPT	0.09	0.09
Adapted CMS	0.06	0.023

In Table 2, results for Enzyme and Parma ham data sets are shown. Theoretically, it does not make sense to impute missing values in these type of patterns for the computation of the dissimilarity measure (See Section 3 for more). As these values do not actually exist, the imputed values can introduce a considerable amount of noise, such that performance of classifiers is affected. However, we applied the factorization methods as imputation on the Autofluorescence data sets to show that this problem holds in practice. The best performance of classifiers is obtained with the modified CMS measure. As explained in Section 3.2, the measure does not take into account the missing values in this case; which is how it should be done according to the nature of the missing data.

5 Conclusions

We have investigated two main approaches with the aim of dealing with the problem of missing values for the classification of multi-way data. The study was

based on the Dissimilarity Representation approach, which consists in building classifiers on a space where objects are represented by their dissimilarities. As a first attempt, factorization techniques were applied to reconstruct the data before dissimilarities were computed. Their performance was good for small and large amounts of missing values, except in the cases where convergence could not be reached. Moreover, they imply an extra computational cost. Therefore, we studied as a second approach, the possibility of computing dissimilarities with the available data only. In this paper, as we experimented on continuous multi-way data, a modification of the Continuous Multi-way Shape measure was introduced in order to deal with missing attributes. This approach, with this particular measure, has shown to be the best option for systematic missing data problems. Moreover, for the other patterns of missing values, it works well when they are present in small amounts(up to 10%). From that point on, classifiers performance deteriorates increasingly. We can then conclude that this approach is suitable for small amounts of missing data. However, the factorization approach should be more reliable for large amounts of missing values. Although experiments were carried out on three-way data only, the presented approaches can be extended to higher-order representations of objects.

References

[1] Mortensen, P.P., Bro, R.: Real time monitoring and chemical profiling of a cultivation process. Chemometr. Intell. Lab. **84**(1-2) (2005) 106–113
[2] Møller, J., Parolari, G., Gabba, L., Christensen, J., Skibsted, L.: Evaluated surface autofluorescence spectroscopy in order to measure age-related quality index of parma ham during processing. J. Agr. Food Chem. **51** (2003) 1224–1230
[3] Tomasi, G., Bro, R.: PARAFAC and missing values. Chemometr. Intell. Lab. **75** (2005) 163–180
[4] Kroonenberg, P.M.: Applied Multiway Data Analysis. John Wiley & Sons. (2008)
[5] Acar, E., Dunlavy, D.M., Kolda, T.G., Mrup, M.: Scalable tensor factorizations for incomplete data. Chemometr. Intell. Lab. **106**(1) (2011) 41–56
[6] Walczak, B., Massart, D.: Dealing with missing data: Part I. Chemometr. Intell. Lab. **58** (2001) 15–27
[7] Smilde, A.K., Bro, R., P., G.: Multi-way Analysis. Applications in the chemical sciences. John Wiley & Sons, Inc. (2004)
[8] Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation For Pattern Recognition. Foundations and Applications. World Scientific. (2005)
[9] Porro-Muñoz, D., Duin, R.P.W., Talavera, I., Orozco-Alzate, M.: Classification of three-way data by the dissimilarity representation. Signal Processing **91**(11) (2011) 2520–2529
[10] Millán-Giraldo, M., Duin, R.P.W., S., S.J.: Dissimilarity-based classification of data with missing attributes. In: Proc. of CIP2010. (2010) 293–298
[11] Porro-Muñoz, D., Duin, R.P.W., Orozco-alzate, M., Talavera, I.: Continuous multi-way shape measure for dissimilarity representation. In: Proc. of CIARP 2012. Volume 7441 of LNCS., Springer (September 2012) 430–437
[12] Lathauwer, L., De Moor, B.: From matrix to tensor: Multilinear algebra and signal processing. In: Proc. of the 4th International Conference on Mathematics in Signal Processing. Volume I., Warwick, UK (1996) 1–11
[13] Gonzalez, R.C., Woods, R.E.: Digital Image Processing (3rd edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2006)
[14] Harshman, R.: Foundations of the Parafac procedure: models and conditions for an explanation multi-modal factor analysis. UCLA Working Papers in Phonetics, Los Angeles **16** (1970) 1–84
[15] Acar, E., Bro, R., Schmidt, B.: New exploratory clustering tool. Journal of Chemomometrics **22** (2008) 91–100