Proceedings of the Federated Conference on Computer Science and Information Systems pp. 245–248
DOI: 10.15439/2018F29
ISSN 2300-5963 ACSIS, Vol. 15

Feature Selection in Time-Series Motion Databases

Florian Elain*, Antonio Mucherino†, Ludovic Hoyet‡, Richard Kulpa§

*INSA, Univ Rennes, Rennes, France. Email: florian.elain@gmail.com
†IRISA, Univ Rennes, Rennes, France. Email: antonio.mucherino@irisa.fr
‡INRIA, Univ Rennes, Rennes, France. Email: ludovic.hoyet@inria.fr
§M2S, Univ Rennes, Rennes, France. Email: richard.kulpa@irisa.fr

***Abstract*—The selection of relevant features in large databases is one of the most important and challenging problems in data mining. Samples forming a given database are generally described by a predefined set of features, and the situation where not all such features can be used for classification purposes needs very often to be faced in real applications. This situation is very typical when the database is related to a phenomenon whose characteristics are not well known. In this context, the extraction of relevant features can therefore also provide additional information on the studied phenomena. We tackle the feature selection problem from an optimization point of view, by reducing it to the problem of finding a maximal consistent "clustering" grouping together the samples and the features of the database. In this work, we extend this approach to dynamical databases, where features are not represented by only one real value, but they are rather given as sequences of a predefined number of real values. Our main contribution consists in proposing an alternative representation of the database so that it fits with a tridimensional matrix with no missing entries, from which a consistent triclustering can be obtained.**

I. Introduction

More and more attention is given nowadays to techniques for mining data, because of the growing amount of information that can be obtained from different resources and that needs to be analyzed [11]. The main aim of such techniques is to identify suitable partitions of a given set of data, where *similar* data can be grouped together. Such partitions can in fact help in finding important relationships in the original data. In some applications, a subset exists for which a classification of the data is already available; the classification associated to this subset can therefore be exploited for *learning* how to classify data for which a classification is not yet known. In this context, our work aims at looking for optimal selections of the features of a dataset in order to improve the quality of the performed classifications.

Let $\mathbb{S}$ be a set of n samples, where every $S_i \in \mathbb{S}$ is represented by an ordered set of m time-series Q_j^i. The number m of time-series per sample is fixed, whereas the length of every time-series can vary. We suppose that a classification of the samples S_i of $\mathbb{S}$, in a given number of classes, is available.

More formally, we suppose that every time-series Q_j^i is a sequence of ℓ_i real values $q_{j,k}^i$, with k counting from 1 to ℓ_i. The length of every time-series depends on the sample S_i, and, since all features of a sample are generally recorded at the same time, we can suppose, without losing generality, that ℓ_i is a constant for all time-series forming the same sample. In brief, we have:

$$\begin{array}{rl} \mathbb{S} = (S_1, S_2, \ldots, S_n) & \text{a set of samples,} \\ S_i = (Q_1^i, Q_2^i, \ldots, Q_m^i) & \text{ordered set of time-series,} \\ Q_j^i = (q_{j,1}^i, q_{j,2}^i, \ldots, q_{j,\ell_i}^i) & \text{time-series.} \end{array}$$

We consider the problem of selecting the subset of time-series that can better describe the phenomena under study. To this purpose, we propose a three-dimensional matrix representation of the original dataset $\mathbb{S}$ that is independent from the length of the time-series, and look for a consistent clustering in sub-matrices where the maximal number of time-series is preserved. Our approach finds its inspiration and extends some previous works (the reader is referred to [7] for a complete description) where non-dynamical problems were considered (every feature was represented by one real value per sample, and not by a time-series).

This short paper is organized as follows. In Section II, we will briefly recall previous works on static problems where the matrix representation of $\mathbb{S}$ is possible with a two-dimensional full matrix. In Section III, we will introduce our three-dimensional matrix representation for datasets where features are represented by time-series. In Section IV, we will propose an extension of the approach recalled in Section II to the data representation introduced in Section III. Finally, Section V will discuss on how to create datasets of human motions to be analyzed by the presented technique, and Section VI will conclude the paper.

II. Feature Selection by Consistent Biclustering

Feature selection is widely studied in the context of data mining. In case the samples of a given dataset do not have a temporal component, the feature selection problem can be tackled by consistent biclustering [7], [9], [10]. This approach works particularly well for problems where measurements are available for every sample, and where the number of features is generally larger than the number of samples in the dataset. The aim, in fact, is to select only important and relevant features from the dataset, whereas others may not be adequate for describing the samples. This gives two immediate consequences. First, if only pertinent features are used and all others are rejected, the memory space necessary for storing the data is optimized. Secondly, a strict relationship between

samples and features can this way be identified, which may reveal important information about the problem under study.

If a set of data contains n samples which are described by m features, then the dataset can be represented by a $m \times n$ matrix $\mathcal{A}$, where the samples are organized column by column, and the features are organized row by row. In this context, we refer to a *bicluster* of $\mathcal{A}$ as a sub-matrix of $\mathcal{A}$, whose elements are a subset of samples and features. Equivalently, a bicluster can be seen as a pair of subsets (S_r, F_r), where S_r is a class (or cluster) of samples, and F_r is a class (or cluster) of features. A *biclustering* [1] is a partition of $\mathcal{A}$ in p biclusters:

$$\mathbb{B} = \{(S_1, F_1), (S_2, F_2), \ldots, (S_p, F_p)\},$$

such that the following conditions are satisfied:

$$\bigcup_{r=1}^{p} S_r \equiv \mathcal{A}, \qquad S_\zeta \cap S_\xi = \emptyset \quad 1 \leq \zeta \neq \xi \leq p,$$

$$\bigcup_{r=1}^{p} F_r \equiv \mathcal{A}, \qquad F_\zeta \cap F_\xi = \emptyset \quad 1 \leq \zeta \neq \xi \leq p,$$

where $p \leq \min(n, m)$ is the number of biclusters.

If a classification for the samples of $\mathcal{A}$ is available, as well as a classification for its features, a biclustering $\mathbb{B}$ can be trivially constructed. Inversely, classifications of samples and features can be obtained from $\mathbb{B}$.

In some data mining applications, there exist sets of data for which a classification of its samples is already given: we say in this case that a *training set* is available. However, the classification of the features used for describing the samples is generally not known, or, equivalently, there is no biclustering $\mathbb{B}$ associated to this training set. Therefore, no a priori information about possible relationships between samples and features is in general given.

A way to obtain a classification for the features from a training set $\mathcal{A}$ is to assign each feature to the class where it is "mostly expressed" (see [7] for a wider discussion). This idea comes from the study of biclusterings related to gene expression data [4], but it can be applied as well to problems arising in other fields (see for example [8]). Once a classification for the features is obtained, a biclustering $\mathbb{B}$ for $\mathcal{A}$ can be computed by simply applying the definition of biclustering. If the found biclustering is *consistent* (in the sense given in [7] for the bidimensional case, the reader is referred to Section IV for additional details), then the selected features are most likely the ones that better describe the samples. In this work, this approach is extended in Section IV to consistent *tri*clusterings.

The feature selection problem can subsequently be formulated as a 0–1 linear fractional optimization problem, which was proved to be NP-hard [5]. We consider a bilevel reformulation of this optimization problem, whose inner problem is linear. For its solution, we employ a heuristic that is based on the meta-heuristic Variable Neighborhood Search (VNS) [3] where, at each iteration, the inner problem is solved exactly.

III. Constructing 3D comparison matrices

As stated in the Introduction, our focus in this work is on datasets whose samples S_i are described by a predefined number of time-series Q^i_j. As in the previous works on consistent biclustering, it is supposed that, for every sample S_i, the same number of time-series Q^i_j are available. Moreover, every pair of time-series Q^A_j and Q^B_j, sharing the same index j but belonging to two different samples, must be related to the same kind of information (e.g. we cannot compare angle variations with the concentration level of a chemical compound). These requirements, which basically ensure that the matrix representation of the biclustering has no missing entries in dimension 2, does not imply a similar property when working with time-series and clustering in 3D. In fact, while the number of samples and the number of features are two constants of the problem (the first two dimensions), the number of elements ℓ_i forming a time-series depends on the sample S_i. Therefore, the corresponding three-dimensional matrix may, in general, have missing entries. Moreover, elements $q^i_{j,k}$ sharing the same index k may have no relationship (whereas common index i means "same sample", and common index j means "same time-series", or equivalently "same feature").

Consider two samples A and B, and two homologous time-series Q^A_j and Q^B_j:

$$(q^A_{j,1}, q^A_{j,2}, \ldots, q^A_{j,\ell_A}), \qquad (q^B_{j,1}, q^B_{j,2}, \ldots, q^B_{j,\ell_B}).$$

In order to obtain a coherent three-dimensional matrix representation, we construct a new matrix where the entries represent comparison scores between pairs of time-series Q^i_j. We consider Dynamic Time Warping (DTW) for a global and temporal alignment of every pair of time-series (see for example [12]). Together with DTW, we also consider the more recent Correlation DTW (CoDTW) [2], which is able to perform better quality alignments in more difficult situations. From the original dataset $\mathbb{S}$, we can therefore compute a full three-dimensional matrix consisting of DTW scores between pairs of samples A and B, for a given feature j:

$$\mathrm{DTW}(A, B; j).$$

A graphical representation of this three-dimensional matrix is given in Fig. 1.

The rows of such a matrix (as well as its columns) contain all (Co)DTW values of one sample S_i in comparison with all the others, for a fixed set of homologous time-series. For this reason, it is reasonable to represent a sample S_i with either a row or a column of such a matrix. This three-dimensional matrix is the result of extending this sample representation to all sets of homologous time-series.

IV. Consistent triclustering

The matrix representation of the original dataset $\mathbb{S}$ that we propose consists of all scores obtained from the time-series comparisons (see previous section). Let DTW be the $n \times n \times m$ matrix containing all such scores. Once the binary vector x is

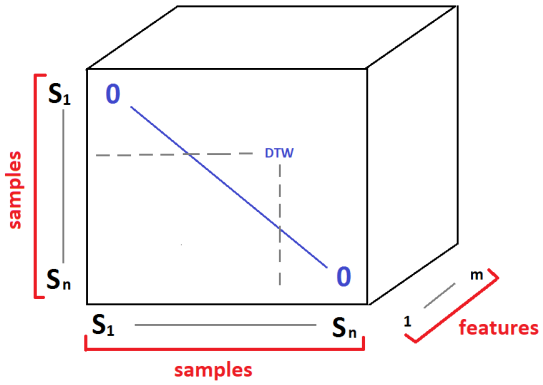

Fig. 1. Score comparison matrix obtained from the original dataset $\mathbb{S}$.

defined over the index set $\{1,2,\ldots,m\}$ so that

$$x_j = \begin{cases} 1 & \text{if the feature } j \text{ is selected} \\ 0 & \text{otherwise,} \end{cases}$$

we can define the sub-matrix DTW[x] obtained by removing from the matrix DTW all features j such that $x_j = 0$.

We suppose that a classification C_S for the samples of $\mathbb{S}$ in p classes is already available. Let $C_S(r)$, with $r \in \{1,2,\ldots,p\}$, indicate the subset of samples belonging to r^{th} class, and let s_{Ar} be a binary parameter indicating whether the sample S_A belongs to the class of samples r. A classification C_F for the homologous sets of time-series can be identified by applying the following rule. For a fixed $\hat{r} \in \{1,2,\ldots,p\}$, the homologous set indexed by $j \in \{1,2,\ldots,m\}$ is assigned to the $\hat{r}^{th}$ class if, and only if, by definition:

$$\forall \xi \in \{1,2,\ldots,p\} \,|\, \xi \neq \hat{r},$$
$$\sum_{A,B \in C_S(\hat{r}) \,|\, A \neq B} \frac{\mathrm{DTW}(A,B;j)}{|C_S(\hat{r})|} < \sum_{A,B \in C_S(\xi) \,|\, A \neq B} \frac{\mathrm{DTW}(A,B;j)}{|C_S(\xi)|}.$$

We suppose working on datasets for which the equation above cannot be satisfied with the equality, otherwise the classification of the features would not be unique.

Let f_{jr} be a binary parameter indicating whether the time-series with index j belongs to the class of features r. A *triclustering* of DTW[x] is *consistent* if

$$\forall \hat{r}, \xi \in \{1,\ldots,p\}, \hat{r} \neq \xi, \forall A,B \in C_S(\hat{r}), A \neq B$$
$$\frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} < \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j}. \quad (1)$$

It is important to remark that the matrix DTW does not admit, in general, a consistent triclustering if all features are considered. Notice that, differently from the previous works, we are interested here in the *less expressed* scores, because they correspond to time-series showing higher similarities.

The problem of selecting the relevant features by consistent triclustering can be stated as follows:

$$\max_x \left(f(x) = \sum_{j=1}^{m} x_j \right)$$
$$\text{subject to} \quad \forall \hat{r}, \xi \in \{1,\ldots,p\}, \hat{r} \neq \xi, \forall A,B \in C_S(\hat{r}), A \neq B$$
$$\frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} < \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j}. \quad (2)$$

As already pointed out, we consider the bilevel reformulation proposed in [7], and we solve the problem by employing a VNS-based heuristic. To perform such a reformulation, we transform the denominators of the optimization problem constraints (see equ.(2)) into continuous variables $y_r, r = 1,\ldots,p$, where y_r represents the number of selected features in the feature class $C_F(r)$:

$$\forall r \in \{1,\ldots,p\}, \; y_r = \sum_{j=1}^{m} f_{jr} x_j.$$

Using the newly introduced variables, the constraint in the original optimization problem can be rewritten by replacing $\sum_{j=1}^{m} f_{j\hat{r}} x_j$ and $\sum_{j=1}^{m} f_{j\xi} x_j$ by $y_{\hat{r}}$ and y_ξ, respectively. We normalize the values:

$$\overline{y}_r = \frac{\sum_{j=1}^{m} f_{jr} x_j}{m},$$

so that the following constraint is satisfied:

$$\sum_{r=1}^{p} \overline{y}_r \leq 1.$$

Our bilevel program is therefore:

$$\textbf{outer pb} \begin{cases} \min_{\overline{y}} \left(g(x,\overline{y}) = \sum_{r=1}^{p} \left[(1 - \overline{y}_r) + \sum_{\xi=1:\xi\neq r}^{p} c(x,r,\xi) \right] \right) \\ \text{subject to} \\ \textbf{inner pb} \begin{cases} x = \arg\max_x \left(f(x) = \sum_{j=1}^{m} x_j \right) \\ \text{subject to consistency constraint (1)} \end{cases} \\ \sum_{r=1}^{p} \overline{y}_r \leq 1, \end{cases}$$

where $c(x,\hat{r},\xi)$ is

$$\sum_{j \in C_S(\hat{r})} \left| \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} - \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j} \right|_+,$$

with $|\cdot|_+$ denoting the function that returns its argument if positive, and 0 otherwise. Hence, $c(x,\hat{r},\xi)$ is strictly positive if and only if at least one constraint is not satisfied.

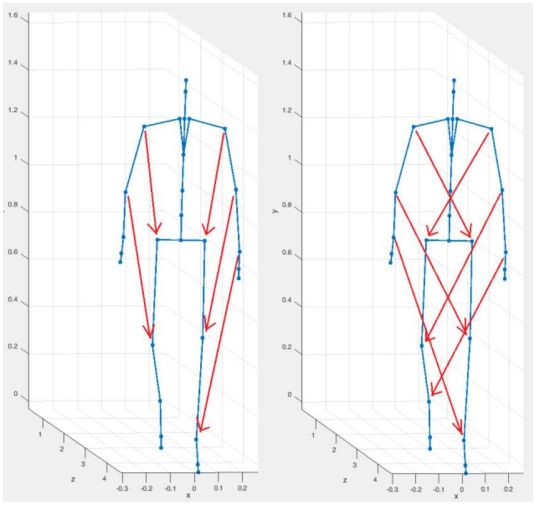

Fig. 2. A graphical representation of some newly introduced features aiming at measuring the symmetries in the movements (see distance sets represented by the red arrows).

V. Creating a Dataset of Human Motions

Motion capture makes it possible to track human movements over time. Markers are generally placed on the body surface of an actor, from which main bones and joint positions of the corresponding skeletal structure, over time, are derived. In general, absolute positions, together with bone rotation angles, are given for every frame of the motion capture (see skeletal representation in Fig. 2, in blue). The data are often stored in a specific format named BVH (BioVision Hierarchy), where joints and bones representing the actor are organized in a hierarchical way.

By collecting a certain number of captured human motions in BVH format, we can define a dataset of motions having particular properties (e.g. all motions are related to a certain human movement, but performed by different classes of humans, such as experts and novices, or male and female). This dataset can serve as a basis for our analyses, but it needs to be manipulated before its effective use.

In fact, the position of each skeleton joint in space may not give very useful information about the movements. Therefore, we propose to enrich the dataset with additional information as follows. For every motion, together with some information related to rotation angles between body parts (which can be easily extracted from BVH files), we also consider relative distances between joint pairs. Some recent studies, in fact, have shown that relative distances can play an important role in the representation of human motions [6]. As Fig. 2 shows, subset of distances can provide information about the symmetry of the movements. All these additional features are represented by time-series.

We have performed some very preliminary experiments where some relevant features were extracted from a so-constructed dataset of human motions (walking motions, with male and female actors) by using our optimization-based approach for feature selection by consistent triclustering. For lack of space, we cannot include any of them in this short paper. As for a future work, we will create a larger collection of motion datasets, having various properties, and we will use them to validate the theory presented in this paper.

VI. Conclusions

We extended an optimization-based approach to feature selection to datasets containing dynamical data. To do so, we proposed an alternative matrix representation of the data, where the original time-series are replaced by similarity scores. This made it possible to extend a previous approach for consistent biclustering to our new three-dimensional matrix representation.

Future works will be aimed at performing supervised classifications by exploiting the information that can be derived from obtained consistent triclusterings. Moreover, we plan to extend this approach to fuzzy sets, so that samples and features can actually belong to more than one class. It is our opinion that this would help better describing real phenomena, such as human motions.

References

[1] S. Busygin, O.A. Prokopyev, P.M. Pardalos, *Feature Selection for Consistent Biclustering via Fractional 0–1 Programming*, Journal of Combinatorial Optimization **10**, 7-21, 2005.

[2] S.A. Etemad, A. Arya, *Correlation-Optimized Time Warping for Motion*, The Visual Computer: International Journal of Computer Graphics **31**(12), 1569–1586, 2015.

[3] P. Hansen and N. Mladenovic, *Variable Neighborhood Search: Principles and Applications*, European Journal of Operational Research **130**(3), 449-467, 2001.

[4] L.-L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z. Weng, G.L. Mutter, M.P. Frosch, M.E. MacDonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, Gr. Stephanopoulos, Ge. Stephanopoulos, S.R. Gullans, *A Compendium of Gene Expression in Normal Human Tissues*, Physiological Genomics **7**, 97-104, 2001.

[5] O.E. Kundakcioglu, P.M. Pardalos, *The Complexity of Feature Selection for Consistent Biclustering*. In: Clustering Challenges in Biological Networks, S. Butenko, P.M. Pardalos, W.A. Chaovalitwongse (Eds.), World Scientific Publishing, 257–266, 2009.

[6] A. Mucherino, D.S. Gonçalves, A. Bernardin, L. Hoyet, F. Multon, *A Distance-Based Approach for Human Posture Simulations*, IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS17), Workshop on Computational Optimization (WCO17), Prague, Czech Republic, 441–444, 2017.

[7] A. Mucherino, L. Liberti, *A VNS-based Heuristic for Feature Selection in Data Mining*. In: "Hybrid Meta-Heuristics", Studies in Computational Intelligence **434**, E-G. Talbi (Ed.), 353–368, 2013.

[8] A. Mucherino, A. Urtubia, *Consistent Biclustering and Applications to Agriculture*, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop on Data Mining in Agriculture (DMA10), Berlin, Germany, 105–113, 2010.

[9] A. Mucherino, P. Papajorgji, P.M. Pardalos, *Data Mining in Agriculture*, 274 pages, Springer, 2009.

[10] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, *A Survey of Data Mining Techniques Applied to Agriculture*, Operational Research: An International Journal **9**(2), 121–140, 2009.

[11] G. Piatetsky-Shapiro, *Advances in Knowledge Discovery and Data Mining*. Usama M. Fayyad, Padhraic Smyth, Ramasamy Uthurusamy (Eds.), vol. 21. Menlo Park: AAAI press, 1996.

[12] X. Xi, E. Keogh, Ch. Shelton, L. Wei, C.A. Ratanamahatana, *Fast Time Series Classification Using Numerosity Reduction*, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 8 pages, 2006.