EuroVis Workshop on Visual Analytics (2015)
E. Bertini and J. C. Roberts (Editors)

Visual Analysis of Relations in Attributed Time-Series Data

Martin Steiger[1], Jürgen Bernard[1], Philipp Schader[1], Jörn Kohlhammer[1,2]

[1]Fraunhofer IGD, Germany
[2]TU Darmstadt, Germany

Figure 1: *Our visual-interactive tool provides an overview of the time series data space (top left). Two complementing views allow for gaining an overview of the categorical attribute space (center and bottom). In a scatterplot view (top right) we link the structures of the time series data (position information) and of the attribute space (similarity-preserving color mapping).*

Abstract

In this paper, we present visual-interactive techniques for revealing relations between two co-existing multivariate feature spaces. Such data is generated, for example, by sensor networks characterized by a set of (categorical) attributes which continuously measure physical quantities over time. A challenging analysis task is the seeking for interesting relations between the time-oriented data and the sensor attributes. Our approach uses visual-interactive analysis to enable analysts to identify correlations between similar time series and similar attributes of the data. It is based on a combination of machine-based encoding of this information in position and color and the human ability to recognize cohesive structures and patterns. In our figures, we illustrate how analysts can identify similarities and anomalies between time series and categorical attributes of metering devices and sensors.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—User-centered design

1. Introduction

The complexity of today's ever-increasing data sources is often assessed in terms of the size, the heterogeneity, or the time-variation. In this work, we focus on the combination of two different feature spaces, e.g., based on time series and multivariate attributes. These types of multi-modal data occur in a variety of application domains such as medical science, manufacturing, and environmental services. Domain experts are challenged with the question of how to gain insight into these types of data in an effective and efficient way. From a high-level perspective domain experts may be interested in the following two questions:

© The Eurographics Association 2015.

DOI: 10.2312/eurova.20151105

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

- Which subsets in the multivariate categorical space are related to a particular time series pattern?
- Which time series patterns are related to a particular subset in the categorical attribute space?

We identify the following challenges for analysis systems providing relation-seeking support for these types of data: Firstly, an overview of the patterns of both individual feature spaces needs to be provided. Secondly, techniques for the combined analysis of *numerical* and *categorical* feature spaces are required. Thirdly, the user needs to be equipped with information drill-down capability in order to focus on specific patterns within the feature spaces. Finally, a mapping concept needs to be implemented allowing for seeking relations in the two co-existing feature spaces.

We contribute a tool which tackles all challenges. On the basis of dimension reduction techniques, our techniques represent the information of two numerical/categorical feature spaces with the visual variables *position* and *color*. The visual variables allow for linking data objects across different views. As a result, analysts are able to relate patterns of one feature space with patterns of the other feature space. We additionally support the relation-seeking task with an interaction design allowing for the selection of meaningful subsets. In different views provided in two figures, we illustrate the techniques in a use case including real-world data of a customer, indicating both efficient and effective relation-seeking support.

2. Related Work

2.1. Dimensionality Reduction

One of the key components of our approach is mapping high-dimensional features spaces into the 2D screen space. Dimensionality reduction techniques are used to create low-dimensional representations of high-dimensional data sets. Based on optimization functions, the algorithms aim at preserving the pairwise distances of data objects. Different distance measures can be used to assess the similarity of (high-dimensional) data objects, see e.g., [BCK08,LRB09] for surveys on distance measures for numerical, categorical, and binary data types. Methods such as the Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) are among the most prominent methods for the reduction of dimensionality [Kru64]. More recently, non-deterministic algorithms such as Stochastic Neighbor Embedding (SNE) methods have been successfully used [HR02]. Joia et al. compare the quality of different approaches with their own projection technique (LAMP) [JPC*11].

2.2. Time Series Feature Extraction

One of our example feature spaces is based on time-oriented data, a data type of specific complexity [AMST11]. Workflows for the extraction of time series feature vectors typically consist of a variety of different steps [DTS*08, Fu11], see [BRG*12, BAF*13] for visual-interactive preprocessing approaches. Typically, workflows include data cleaning steps with respect to quality levels, specific tasks or user needs [GGAM12]. The application of a time series descriptor transforms the time series into a compact and yet representative feature space [KK03]. In this work, we apply the Perceptual Important Points algorithm (PIP) to preserve the shape of the time series, similar to the approach by Ziegler at al. [ZJGK10]. We refer to the works of Andrienko and Andrienko [AA06] and Aigner et al. [AMST11] for an overview of various visual representations of time-oriented data.

2.3. Comparison and Relation-Seeking Approaches

With our technique, we focus on the analysis tasks of *comparing* patterns within feature spaces and *seeking for relations* in between feature spaces [AA06]. For relation and comparison techniques, e.g., for multi-modal data, we refer to the survey of Kehrer and Hauser [KH13]. Gleicher et al. [GAW*11] review approaches facilitating the visual comparison of data. For time series data most similar is the work by Steiger et al. where the output of a time series projection can be compared at a synoptic level [SBM*14]. However, while Steiger et al. relate the time series data to geospatial network, our approach reveals relations between the time series data content and an additional multivariate attribute space. An approach where single metadata attributes are related to the content of time series data is presented by Bernard et al. [BRS*12]. The approach is inspiring with respect to the mosaic metaphor revealing the distribution of data elements. However, the technique does not allow for the analysis of multiple attributes. Considering mixed data sets relation seeking approaches are, e.g., Parallel Sets [KBH06], the contingency wheel [AAMG12], and a technique emphasizing mixed research data [BSW*14].

3. Data & Algorithms

3.1. Characterization of Supported Data Types

We focus on multi-modal data sets consisting of two separate multivariate feature spaces. These spaces can either be completely of numerical or of categorical data type. The real-world data in our use case is a collection of time series data linked to a set of mixed data attributes. We apply binning techniques to discretize continuous attributes of the mixed data set. In this manner, we create a multivariate categorical data set, each attribute containing a number of different *bins*. Missing values in the categorical attributes are treated as an additional bin. We use time series data since it is a relevant representative, allowing for the extraction of multivariate numerical features and for the visual representation of both raw data and features. Our technique is independent of specific workflows for the extraction of time series features. The workflow carried out in our use case uses the Perceptual Important Points (PIP) algorithm [ZJGK10] as a shape-preserving time series descriptor invariant to missing values.

© The Eurographics Association 2015.

3.2. Mapping Multivariate Data into 2D Space

A variety of projection algorithms exist allowing for the calculation of low-dimensional representations of feature sets. We focus on the class of MDS projections that take pairwise object distances as an input. In this way, our mapping supports data objects of both numerical and categorical data type. The data type determines the class of distance measure. For the time series data resulting in a numerical feature space, we use the Dynamic Time Warping (DTW) algorithm as described by Salvador and Chan [SC07]. The DTW is a flexible distance measure with respect to small distortions and therefore well suited for the PIP features of our use case.

For the categorical data, we apply the Hamming distance to identify the number of attributes that differ with respect to the bin distribution. Equal bins of categorical attributes have distance 0, different bins have distance 1. For ordinal (i.e. ranked) attributes this distance measure is still discrete, but more detailed (0, 1, 2, etc).

3.3. Linking The Results of Two 2D Projections

With our projection techniques, we receive two 2D representations of the data set. Both representations preserve the structure of an individual feature space. We now present the functional support for our visual-interactive linking solution, which is presented in Section 4.

Our solution is inspired by the results of an experiment performed by Ware and Beatty [WB88]. Its results indicate that encoding information in color can be as powerful as the positional encoding. Thus, we use the visual variables *position* and *color* to encode the structural information of two individual feature spaces. The position information of the first projection output is direcly mapped to the visual variable position. The position information of the second projection is mapped onto a static 2D colormap. The 2D colormap transforms the position information in the visual variable color. A variety of 2D colormaps exists that faithfully represents geometric similarity in color similarity (called perceptual linearity). The work of Bernard et al. gives an overview on different 2D colormaps and also provides task-based recommendations [BSM*15] for the choice of 2D colormaps. We chose the colormap by Teuling et al. since it provides both a large amount of distinguishable colors and a particularly good perceptual linearity [TSS11]. The result of the linking step is a point plot, where the position and the color information reveal the structure of two multivariate feature spaces.

4. Visualization & Interaction

We present the visualization and interaction design of our analysis tool. On the basis of the data abstraction and the functional support presented in Section 3, the tool provides an overview of both feature spaces, drill-down capability, and a linking concept across different views. Four views are depicted in Figure 1: the *Macro View* (top left), the *Micro View* (top right), the *Attribute Table* (center), and the *Bin Mosaic Plot* (bottom).

4.1. Time Series: Micro View and Macro View

We start with an overview of the time series feature space. The two projection views show the projected time series data on different levels of abstraction, inspired by Tufte's concept of micro-macro views [Tuf90] [BvLBS09]. The *Micro View* renders all elements with a scatterplot metaphor, while the *Macro View* shows the results of a spatial aggregation of the projected data by means of a glyph.

In the *Micro View*, every time series is represented with a single colored dot. The *position* information of the data elements stems from the time series projection. The *color* information of the data elements is defined by the projection of the categorical attributes (cf. Section 3.3). Both the position and the color information of objects allows for the identification and the comparison of patterns. In addition, these patterns enable analysts to seek for relations across both feature spaces. On demand, tooltip functionality allows for the identification of both the time series and the attribute information of individual objects.

The *Macro View* shows the information of the *Micro View*, enhanced by an additional data abstraction layer. For that purpose, we provide a data aggregation step which is visually represented as line charts on top of the scattered data. Our implementation is based on the k-means clustering algorithm for the aggregation of time series features. Since we apply the algorithm for providing an overview of the time series space, the number of k does not necessarily need to comply with the number of 'intrinsic' data clusters of the data set. It is rather important to visually represent the structure of the high-dimensional data space.

4.2. Attributes: Attribute Table and Bin Mosaic Plot

Two complementing views enable the analyst to access the attribute space.

The *Attribute Table* (top) is composed of a table layout where every attribute corresponds to a column and every bin corresponds to a cell, respectively. Labels are used to (a) depict attribute names, (b) bin names, and (c) the number of objects in every bin. The relative amount of selected elements of a bin is represented by a gauge bar metaphor. The colors of the gauge bars match the attribute projection coloring presented in the Micro View. Sorting the segments by hue allows for the comparison of cell elements and cells, respectively. Technically speaking, these are stacked bar charts for each attribute on a relative scale. In Figure 1, only a subset of the data was selected, leading to partially filled bars.

The *Bin Mosaic Plot* complements the Attribute Table by showing the distribution of the attribute feature space from

© The Eurographics Association 2015.

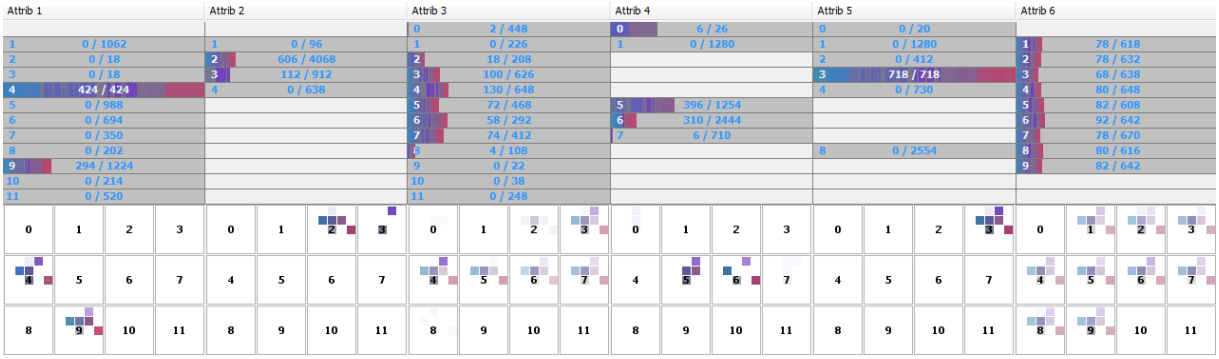

Figure 2: *The two Attribute Views: in the Attribute Table (top), six attributes are enumerated from left to right. Each bin has its own cell representing a respective data subset. The Bin Mosaic Plot (bottom) plots the bins as projected in the colormap.*

a different perspective. For that purpose, we utilize the position information of the attribute projection onto the color map (cf. Section 3.3). A mosaic metaphor represents the data distribution of the attribute projection, discretized to a 2D grid. The static color map assigns colors to each of the mosaic tile. We use transparency value to indicate the absolute number of data elements per block. As a result, the Bin Mosaic Plot complements the Attribute Table by showing the absolute number of elements for each bin. For example, the Bin Mosaic Plot highlights that the bins in 'Attribute 2' in Figure 1 strongly differ in size. In addition, the 2D arrangement of the Bin Mosaic Plot facilitates the comparison of different color distributions of individual bins. For example, the 'Attribute 5' in Figure 1 consists of heterogeneous bin distributions, indicated by varying mosaic tiles for each bin.

4.3. Interaction

We provide two directions of interaction for seeking relations in the two feature spaces: time- and attribute-based.

We start with time-based interaction: The user can select an arbitrary group of time series elements in the *Micro View* using a lasso selection tool. This can be done, for example, to select groups that contain similarly colored points. As an alternative, a time series aggregate can be selected in the *Macro View*. Since the views are linked, the same elements are also selected in the other views as well. Consequently, the attribute views display the distribution of the bins of the selected elements. In this way, the user is enabled to relate a data subset of a particular time series content with the attribute space. An interesting time-based interaction can be seen in the use case illustrated in Figure 1. With the lasso tool a set of proximate time series elements was selected leading to complete different selected subsets in the bin visualizations. As an example, in both Attribute 1 and Attribute 2 two bins are almost completely selected. This brings two insights. Firstly, each of the attributes have a strong relation to the time-based selection. Secondly, the two attributes seem to be correlated, at least for the currently selected subset. In contrast, Attribute 6 seems to be rather unaffected.

The second type of interaction is based on bins. Clicking on a cell selects all corresponding data elements, all other views are updated, reflecting the selection. All un-selected elements are dimmed in the *Micro View*. With the selection of a bin, the user is able to (a) identify and compare the distributions of the data subset in the remaining attributes and (b) identify the distribution of the data subset in the time series projections. The latter also enables the user to detect patterns or cohesive structures in the time series plot. In this manner, the analyst can identify interesting relations between the attribute space and time series space. The result of a bin selection is illustrated in Figure 2. In the corresponding use case, it was possible to identify that the bin selection is highly related to time series patterns at a certain area of the scatterplot. The Macro View (left) allows for an enhanced look-up of the value distribution of the data elements at the center of the projection. The selected bin is highly related to time series patterns with a value domain fluctuating around zero.

5. Conclusion

We presented techniques allowing for revealing relations between two co-existing multivariate feature spaces for a set of data elements. The techniques are presented in a Visual Analytics tool with an example time series data set with mixed attributes. The tool allows for the analysis of both time series and multi-variate data. As a result, the analyst is enabled to identify interesting relations between the two spaces. The illustrated results show that the tool allows analysts to carry out comparison and relation seeking tasks.

While the focus of this work is on the provided techniques, we are using the tool in two case studies with real users. We identify the need for more user parameters for the feature extraction pipeline, the choice of the projection methods, and for the visual representation of the high-dimensional features.

© The Eurographics Association 2015.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 2

[AAMG12] ALSALLAKH B., AIGNER W., MIKSCH S., GROLLER M.: Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2849–2858. 2

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*. Springer, London, UK, 2011. doi:10.1007/978-0-85729-079-3. 2

[BAF*13] BÖGL M., AIGNER W., FILZMOSER P., LAMMARSCH T., MIKSCH S., RIND A.: Visual analytics for model selection in time series analysis. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2237–2246. 2

[BCK08] BORIAH S., CHANDOLA V., KUMAR V.: Similarity measures for categorical data: A comparative evaluation. *In Proceedings of the 8th SIAM International Conference on Data Mining 30*, 2 (2008), 3. 2

[BRG*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-interactive preprocessing of time series data. In *SIGRAD* (2012), pp. 39–48. 2

[BRS*12] BERNARD J., RUPPERT T., SCHERER M., KOHLHAMMER J., SCHRECK T.: Content-based layouts for exploratory metadata search in scientific research data. In *Proc. of the Joint Conference on Digital Libraries* (2012), JCDL, ACM, pp. 139–148. doi:10.1145/2232817.2232844. 2

[BSM*15] BERNARD J., STEIGER M., MITTELSTÃĎDT S., THUM S., KEIM D., KOHLHAMMER J.: A survey and task-based quality assessment of static 2D colormaps. *Proc. SPIE 9397* (2015), 93970M–93970M–16. URL: http://dx.doi.org/10.1117/12.2079841, doi:10.1117/12.2079841. 3

[BSW*14] BERNARD J., STEIGER M., WIDMER S., LÃIJCKE-TIEKE H., MAY T., KOHLHAMMER J.: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum 33*, 3 (2014), 291–300. 2

[BvLBS09] BERNARD J., VON LANDESBERGER T., BREMM S., SCHRECK T.: Micro-macro views for visual trajectory cluster analysis. In *IEEE Symposium on Visualization* (2009). 3

[DTS*08] DING H., TRAJCEVSKI G., SCHEUERMANN P., WANG X., KEOGH E.: Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow. 1*, 2 (Aug. 2008), 1542–1552. 2

[Fu11] FU T.-C.: A review on time series data mining. *Eng. Appl. Artif. Intell. 24*, 1 (Feb. 2011), 164–181. doi:10.1016/j.engappai.2010.09.007. 2

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (2011), 289–309. doi:10.1177/1473871611416549. 2

[GGAM12] GSCHWANDTNER T., GÄRTNER J., AIGNER W., MIKSCH S.: A taxonomy of dirty time-oriented data. In *CD-ARES* (2012), Quirchmayr G., Basl J., You I., Xu L., Weippl E., (Eds.), vol. 7465 of *Lecture Notes in Computer Science*, Springer, pp. 58–72. 2

[HR02] HINTON G. E., ROWEIS S. T.: Stochastic neighbor embedding. In *Advances in neural information processing systems* (2002), pp. 833–840. 2

[JPC*11] JOIA P., PAULOVICH F. V., COIMBRA D., CUMINATO J. A., NONATO L. G.: Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2563–2571. 2

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568. 2

[KH13] KEHRER J., HAUSER H.: Visualization and visual analysis of multifaceted scientific data: A survey. *Visualization and Computer Graphics, IEEE Transactions on 19*, 3 (2013), 495–513. doi:10.1109/TVCG.2012.110. 2

[KK03] KEOGH E., KASETTY S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min. Knowl. Discov. 7*, 4 (2003), 349–371. doi:10.1023/A:1024988512476. 2

[Kru64] KRUSKAL J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27. doi:10.1007/BF02289565. 2

[LRB09] LESOT M., RIFQI M., BENHADDA H.: Similarity measures for binary and numerical data: a survey. *Int. J. Knowl. Eng. Soft Data Paradigm. 1*, 1 (Dec. 2009), 63–84. URL: http://dx.doi.org/10.1504/IJKESDP.2009.021985, doi:10.1504/IJKESDP.2009.021985. 2

[SBM*14] STEIGER M., BERNARD J., MITTELSTÄDT S., LÜCKE-TIEKE H., KEIM D., MAY T., KOHLHAMMER J.: Visual analysis of time-series similarities for anomaly detection in sensor networks. *Computer Graphics Forum 33*, 3 (2014), 401–410. doi:10.1111/cgf.12396. 2

[SC07] SALVADOR S., CHAN P.: Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal. 11*, 5 (2007), 561–580. 3

[TSS11] TEULING A. J., STÖCKLI R., SENEVIRATNE S. I.: Bivariate colour maps for visualizing climate data. *International Journal of Climatology 31*, 9 (2011), 1408ï£¡1412. doi:10.1002/joc.2153. 3

[Tuf90] TUFTE E.: *Envisioning Information*. Graphics Press, Cheshire, CT, USA, 1990. 3

[WB88] WARE C., BEATTY J. C.: Using color dimensions to display data dimensions. *Hum. Factors 30*, 2 (Apr. 1988), 127–142. URL: http://dl.acm.org/citation.cfm?id=46356.46357. 3

[ZJGK10] ZIEGLER H., JENNY M., GRUSE T., KEIM D.: Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), pp. 83–90. doi:10.1109/VAST.2010.5652530. 2

© The Eurographics Association 2015.