Visualization of Trends Using RadViz

Lenka Nováková and Olga Štěpánková

Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic
{novakova,step}@labe.felk.cvut.cz

Abstract. Data mining is sometimes treating data consisting of items representing measurements of a single property taken in different time points. In this case data can be understood as a time series of one feature. It is no exception when the clue for evaluation of such data is related to their development trends as observed in several successive time points.

From the qualitative point of view one can distinguish 3 basic types of behavior between two neighboring time points: the value of the feature is stable (remains the same), it grows or it falls. This paper is concerned with identification of typical qualitative development patterns as they appear in the windows of given length in the considered time-stamped data and their utilization for specification of interesting subgroups.

Keywords: Time Series, Data Visualization, RadViz.

1 Introduction

The paper explains a novel approach to the search of typical qualitative patterns of development trends as they appear in windows of time stamped sequences of measurements of a single feature (e.g. weight of a person). Our solution is based on RadViz [2] 2D visualization of relevant data which is implemented in our SW tool. The applicability of the suggested solution is tested and demonstrated on the Stulong data set.

Our approach takes advantage of an interesting feature of RadViz visualization method [3] [4], namely its ability to depict some relational dependencies as studied in [7]. The Fig. 1 offers a clear example of a relational dependency among data that can be easily identified in a RadViz picture. This is a natural consequence of the RadViz mapping definition as briefly explained in the Section 2. Specific RadViz properties and modifications useful for analysis of time stamped data as well as design principles of our SW tool are treated in the Section 3. Some results obtained for Stulong data are reviewed in the Section 4. We conclude by pointing to several open questions in the last section where our plans for further development are outlined.

J. Rauch et al. (Eds.): ISMIS 2009, LNAI 5722, pp. 56–65, 2009.
© Springer-Verlag Berlin Heidelberg 2009

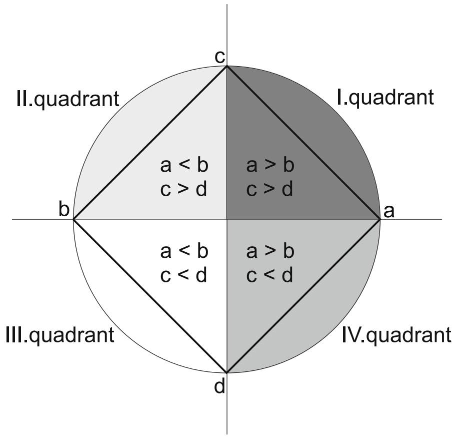


Fig. 1. The RadViz can show relations

2 Description of RadViz Mapping

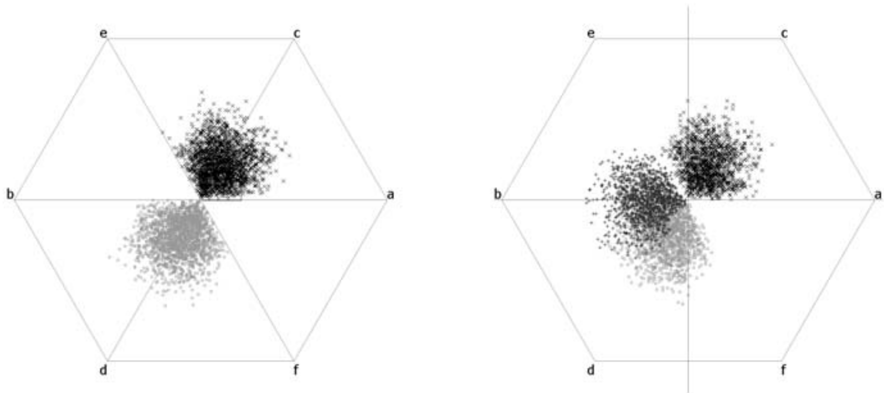
RadViz is a mapping from n-dimensional space into a plane [2] [3] [4]. The point $y = [y_1, \ldots, y_n]$ in an n-dimensional space is mapped to the point $u = [u_1, u_2]$ with following coordinates:

$$u_1 = \frac{\sum_{j=1}^{n} y_j \cos(\alpha_j)}{\sum_{j=1}^{n} y_j}$$
$$u_2 = \frac{\sum_{j=1}^{n} y_j \sin(\alpha_j)}{\sum_{j=1}^{n} y_j}$$

Let us consider the case depicted on the Fig. 1. Here, the angles corresponding to the anchors a, b,c and d are $\alpha = 0$, $\beta = \pi$, $\gamma = \frac{\pi}{2}$ and $\delta = \frac{3\pi}{2}$. After substituting the respective values into these equations we get the simplified equation for the point $u = [u_1, u_2]$:

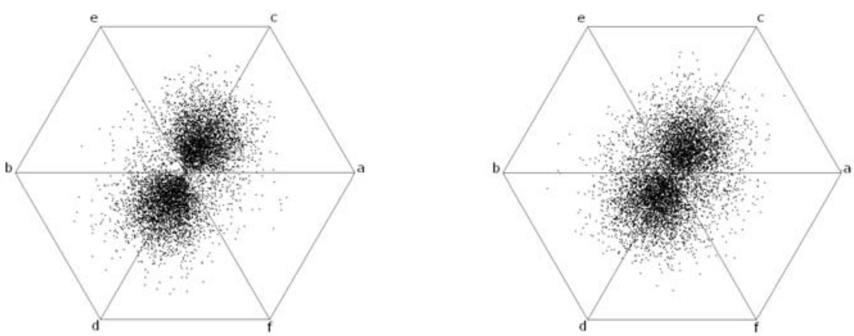
$$u_1 = \frac{a - b}{a + b + c + d}$$
$$u_2 = \frac{c - d}{a + b + c + d}$$

All records from the data set, where the value of the attribute a is greater than that of b, are depicted in the first and the fourth quadrant of RadViz graph, see the Fig. 1. The data are divided by the vertical axis, which runs between anchors a and b and crosses the origin. On its right side there lie all points, where a is greater than b. On its left side, there lie the points where a is less than b. On the axis itself there are all the points for which a equals b. The



(a) (b)

Fig. 2. Visualization of a relational dependency: a) $a < b$ & $c < d$ & $e < f$ in square points, $a > b$ & $c > d$ & $e > f$ in cross points color b) data from previous the image joined with those meeting the constraint ($a < b$ & $c < d$ & $e > f$)



(a) (b)

Fig. 3. Relation between three pairs: a) 10% noisy b) 30% noisy

same situation appears in the case of the attributes c and d. The Fig. 1 shows that given two pairs of attributes we can clearly distinguish those sections of the RadViz picture corresponding to various possible conjunctive combinations of their relations ($a < b$ & $c < d$, etc.). Each of the 9 possible non-overlapping sets of data described by upper mentioned constraints is mapped into a separate subset of the RadViz image (these sets represent its disjunctive partition).

Unfortunately, there is no straightforward generalization of the situation observed in the Fig. 1 as soon as there are considered sets of attribute pairs of cardinality higher than 2. Even in this case we can identify sections in the RadViz image corresponding to specific patterns of qualitative behavior described by conjunction of relations between the values of pairs of the considered attributes, see for example Fig. 2.a.

We have to be aware of fact that the section corresponding to the data points meeting the condition $(a < b \;\&\; c < d \;\&\; e > f)$ has a large overlap with the section $(a < b \;\&\; c < d \;\&\; e < f)$. The images of the disjoint data subsets of our interest are no more disjoint and this causes problems in the interpretation of a RadViz image, see the Fig. 2.b - denote that the 3 depicted sets are of the same size. The next two figures Fig. 3.a and Fig. 3.b show the situation from Fig. 2.a with 10% and 30% of noise, respectively.

3 Visualization of Time Series in RadViz

Time series is a sequence of samples measured in fixed time points. Often, the data result from repeated measurements of a certain feature (e.g. blood pressure) for individual studied objects (e.g. patients). For simplicity let us assume that all the measured values appear in the interval $\langle 0, 1 \rangle$. If this is not the case the values have to be normalized using the global minimum and maximum for all the used attributes. Whenever the values of two attributes are the same they remain the same after normalization, of course. Such data are frequently visualized by a graph called "parallel coordinates", where values corresponding to a single object are connected by a single broken line, see the Fig. 4.

Description of data through parallel coordinates provides valuable information if the size of the data set is small and the lines corresponding to individual objects can be easily distinguished e.g. by different colors. But as soon as the number of studied objects grows, the picture created in parallel coordinates provides no more evidence about the studied data. In general, one cannot see from the picture whether there is any dominant pattern appearing in the data or not, see the Fig. 9. Let us search for the qualitative development patterns appearing in the studied data. The value between two neighboring attributes can "grow", "be stable" or "fall". If we have measurements in $n+1$ consecutive time points, we can observe 3^n different shapes of the corresponding time series. In the case where we distinguish only behavior types "grow" or "fall" there exist 2^n shapes. Clearly, systematic search through such a space does not seem reasonable. Fortunately, it is not necessary to solve the task of finding large enough interesting subsets of time series data that are characterized by a single qualitative development trend (e.g. "grow", "fall", "grow"):

Suppose, we are studying data series of the length m described by attributes $[at_1, \ldots, at_m]$ with the aim to find interesting subsets of data that are characterized by a qualitative development trend of the length lower than m and starting from the value of at_1. If the lower bound for the cardinality of the target set is given we can apply Apriori-like approach and start by the search for the interesting subsets characterized by the trends among the initial three attributes $[at_1, at_2, at_3]$. Any interesting subset has to be a refinement (or more precisely

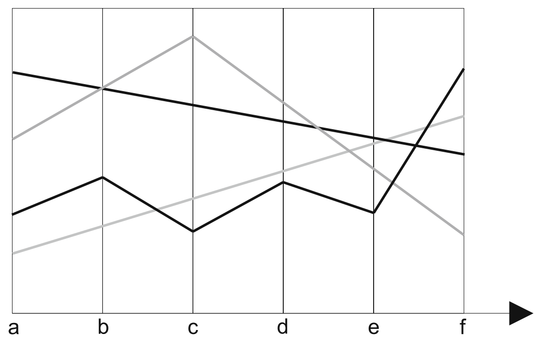


Fig. 4. Illustration example of time series

an extension towards further attributes) of an interesting subset which has been identified in the former step. In other words, we can focus our attention to the sets identified in the first step only and proceed by searching for their refinement using $at_4, , at_3$ etc. recursively.

We already know that RadViz can easily depict the relation "less than". Let us utilize this feature to visualize time series. Let us denote the values corresponding to measurements in the time points 0, 1, 2 and 3 by attributes a, b, c and d. The trend between the attributes a and b from Fig. 4 is $b - a$ and its qualitative value can be naturally identified from its position in the RadViz picture without any need for additional computation provided the anchors a and b are situated opposite each other as in the Fig. 1.

Let us consider two pairs of successive attributes, the first pair is a,b and the second pair is b,c and a RadViz image with 4 dimensional anchors, where two anchors are bound to the attribute b. The corresponding anchors are denoted as b_1 and b_2) in the Fig. 5. We can replace this doubled dimensional anchor by a single dimensional anchor as shown in the the Fig. 5. The RadViz projections resulting from both the anchor settings are the same - compare for example the RadViz images at the Fig. 7.a and the Fig. 7.b of the Stulong data set treated in the following section with 4 and 3 dimensional anchors.

In the rest of the paper, we will work with the specific RadViz projections using three anchors only, that are identical to RadViz projections with two pairs of anchors. The radius of the dimensional anchor a is 1 and the angle $\alpha = 0$. The dimensional anchor for attribute b has radius $\sqrt{2}$ and angle is $\beta = \frac{3}{4}\pi$. This anchor lies further from center than the anchor for attributes a and c. The dimensional anchor for attribute c has radius 1 and its angle is $\gamma = \frac{3}{2}\pi$.

The equation for counting point $u = [u_1, u_2]$ of RadViz image are:

$$u_1 = \frac{a - b}{a + b + c}, \quad u_2 = \frac{b - c}{a + b + c}. \tag{1}$$

The design of anchors and the key how to interpret the trends as they appear in the RadViz image is shown in the Fig. 5.

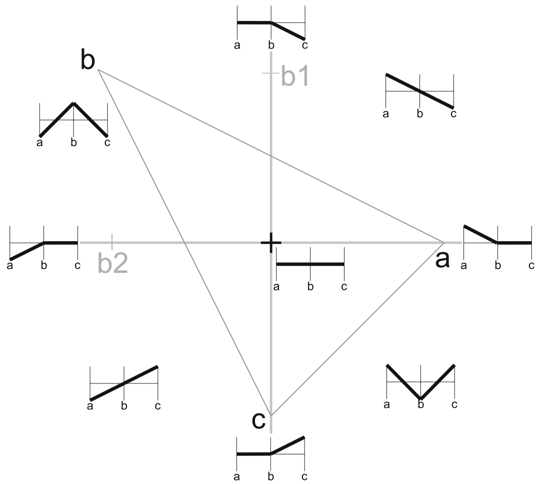


Fig. 5. How to detect trends in RadViz

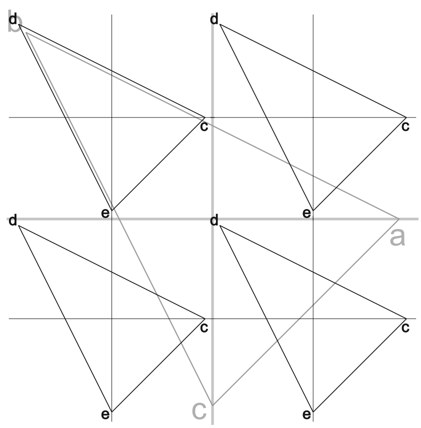


Fig. 6. Recursive expansion of RadViz image

Obviously, information about the development trends in the intervals from a to b and from b to c can be depicted in the orthogonal coordinates representing the values $(b-a)$ and $(c-b)$, too. We have decided to use RadViz visualization because it works directly with the data as they are (no need to create a new matrix with differences of neighboring attributes) and our SW tool implementing modification of RadViz offers number of additional features that proved helpful in the phase of data understanding and we can take its advantage e.g. when searching for frequent development trends.

This usage of RadViz can complement other visualization methods for example Parallel Coordinates as it can help the user to select group of data records with some trends. This is impossible to be done directly in Parallel Coordinates. Eventuality, the RadViz graph can be expanded recursively as is suggested in the Fig. 6.

4 Identification of Trends in Stulong Data Set

The Stulong data set[1] comes from a truly longitudinal study aimed at primary prevention of atherosclerotic cardiovascular diseases (CVD). The study covers observation of more than 1400 middle aged men during 20 years. The intention of the project was to identify significant atherosclerosis risk factors, to follow development of these risk factors and their impact on the examined men health, especially with respect to atherosclerotic cardiovascular diseases.

Let us try to identify difference in time development of monitored physiological attributes between the two groups of patients classified with respect to occurrence of cardiovascular disease:

- patients who fell in by cardiovascular disease during the study (CVD = 1)
- patients who remained healthy during the study (CVD = 0)

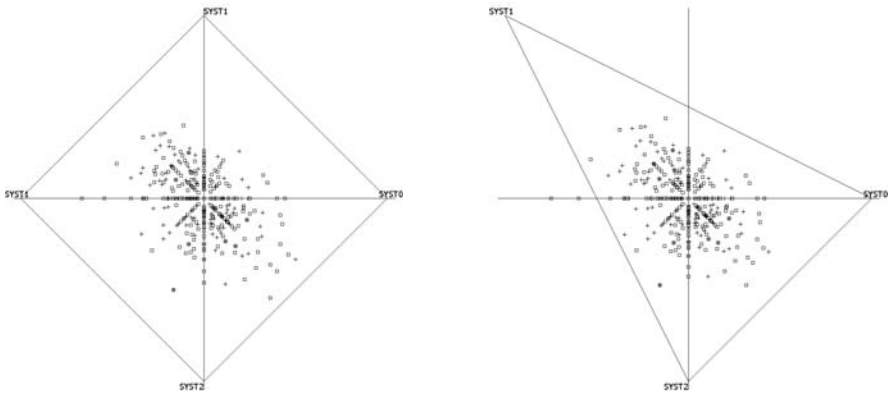
The data from Stulong data set was preprocessed and there was proposed a new definition of derived attributes by combination of windowing and discretization [6]. These derived attributes were used for interesting sub-group discovery using association rules [9].

The RadViz method can help in the data understanding phase by depicting the first view to the trends appearing in the data. The Fig. 7.b shows the trends of systolic blood pressure in two neighboring intervals (between Syst0, Syst1 and between Syst1, Syst2). The CVD classification of the studied data offers an additional means for specification of an interesting subgroup. It is such a subgroup where the frequency of the classes is significantly different than that in the original set of data.

Our implementation of RadViz enables to explore the graph in detail. The user can interactively select by a rectangle the group of points and the system returns the table of all corresponding data records. Moreover, it summarizes the relevant information using a pie graph: while the inner pie gives the distribution of data in the full (original) data set, the outer pie describes the selected group, see the Fig. 8.a and Fig. 8.b.

The Fig. 8.a and Fig. 8.b show the difference between the whole set of patients and the selected group of the patients. The group of patients, who fall in

[1] The STULONG study was performed at the 2nd Dept. of Internal Medicine of the 1st Medical Faculty of the Charles University and General Faculty Hospital in Prague 2 (initially General Medical Faculty of the Charles University and Faculty Hospital I) managed by clinicians prof. MUDr. F.Boudik, DrSc., MUDr. M.Tomeckova, CSc. and doc. MUDr. J.Bultas, CSc. Most of the data was transferred into the electronic form by EuroMISE (European Centre for Medical Informatics, Statistics and Epidemiology) with support of European project [10].



(a) (b)

Fig. 7. Visualization of trends between two interval of systolic blood pressure (between Syst0, Syst1 and between Syst1, Syst2)

cardiovascular disease, are depicted in light grey color. The group of patients, who stay healthy, is depicted in dark grey color.

The difference between the whole group and patients with falling systolic blood pressure in both intervals are shown in the Fig. 8.a. This group fell in by cardiovascular disease less often than the whole group. On the other hand the Fig. 8.b shows the difference between the whole set and patients with growing systolic blood pressure in both intervals. This group has more than 1.5 times frequency that they fall in by cardiovascular disease. It is obvious that there is the difference between these two types of trends.

Using the suggested approach we were able to find quickly more interesting subgroups than those reported in [8]. One of such interesting subgroups is that

(a) (b)

Fig. 8. Difference between CVD occurrences in the original data set and a subset specified through qualitative development trends: a) ($Syst0 > Syst1$ & $Syst1 > Syst2$), ie. the trend of systolic blood pressure is “fall”-“fall” and b) ($Syst0 < Syst1$ & $Syst1 < Syst2$), i.e. the trend of systolic blood pressure is “grow”-“grow”

of patients, who have in the beginning stable diastolic blood pressure $Diast0 = Diast1$ and $Diast1 = Diast2$ followed by a jump up, i.e. $Diast2 < Diast3$. The group has 21 members (it is 5% of whole population) and there is a significant difference in the frequency of the patients with CVD=0 in this subgroup when compared to the original set, namely 27%.

5 Conclusion

The RadViz method provided significant help in search for qualitative patterns in the trends in the Stulong data set. Our experience proves that the RadViz visualization complements well the Parallel Coordinates [5] approach, see the Fig. 9: while Parallel Coordinates provide the global view to data, RadViz image can mediate more accurate local view.

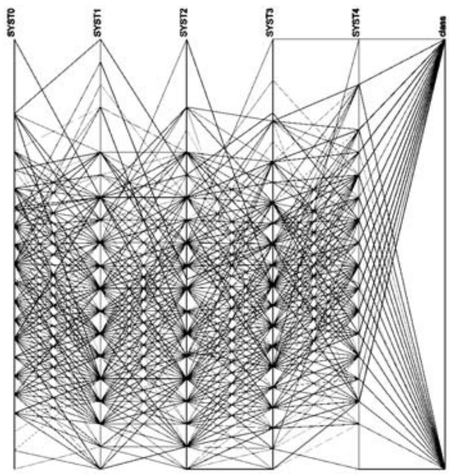


Fig. 9. Visualization of Stulong data set by Parallel Coordinates

Currently, we are finishing an implementation of a SW tool for data analysis that is extensively utilizing RadViz visualization. One of services it will offer is the systematic top-down automated search for frequently occurring qualitative trends. The tool applies the upper mentioned solution inspired by Apriori algorithm [1] to stop the recursive expansion, whenever there is not enough data in the expanded groups. Of course, it will ensure frequently used statistical evaluation to verify significance of the identified subsets. For practical applications it might be useful to introduce a possibility to work with fuzzy relations (e.g. “fuzzy equal”). This will be considered in further releases.

Acknowledgements. The presented research and development has been supported by the grant 1ET101210513 (Relational Machine Learning for Biomedical Data Analysis).

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast Discovery of Association Rules. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, ch. 12, pp. 307–328. AAAI/MIT Press, Cambridge (1996)
2. Fayyad, U.M., Grinstein, G.G., Wierse, A.: Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, San Francisco (2002)
3. Hoffman, P., Grinstein, G., Marx, K., Grosse, I., Stanley, E.: DNA Visual And Analytic Data Mining. In: Proceedings of the IEEE Visualization 1997 Conference, pp. 437–441 (1997)
4. Hoffman, P., Grinstein, G., Pinkney, D.: Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In: Proceedings of the 1999 Workshop on New Paradigms in information Visualization and Manipulation, pp. 9–16 (1999)
5. Inselberg, A.: The Plane with Parallel Coordinates. Special Issue on Computational Geometry, The Visual Computer 1, 69–91 (1985)
6. Kléma, J., Nováková, L., Karel, F., Štěpánková, O., Železný, F.: Sequential Data Mining: A Comparative Case Study in Development of Atherosclerosis Risk Factors. IEEE Transactions on Systems, Man, and Cybernetics: Part C 38(1), 3–15 (2008)
7. Nováková, L., Štěpánková, O.: Visualization of Some Relational Patterns for DM. In: Cybernetics and Systems 2006, Vienna, vol. 2, pp. 785–790 (2006)
8. Nováková, L., Karel, F., Aubrecht, P., Tomečková, M., Rauch, J., et al.: Trends in time windows as risk factors of cardiovascular disease. In: Znalosti 2008, pp. 148–159. Slovak University of Technology, Bratislava (2008) (in Czech)
9. Rauch, J., Šimůnek, M.: GUHA Method and Granular Computing. In: Hu, X., et al. (eds.) Proceedings of IEEE conference Granular Computing, pp. 630–635 (2005)
10. Project STULONG, WWW page, `http://euromise.vse.cz/stulong`