Qualitative Clustering of Short Time-Series: A Case Study of Firms Reputation Data

Ljupčo Todorovski[1], Bojan Cestnik[2], Mihael Kline[3], Nada Lavrač[1], Sašo Džeroski[1]

[1] Department of Intelligent Systems, Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
{Ljupco.Todorovski, Nada.Lavrac, Saso.Dzeroski}@ijs.si
[2] Temida, Grassellijeva 20, SI-1000 Ljubljana, Slovenia
Bojan.Cestnik@temida.si
University of Ljubljana, Faculty of Social Sciences,
Kardeljeva pl. 5, SI-1000 Ljubljana, Slovenia
Mihael.Kline@uni-lj.si

Abstract

In this paper, we propose a clustering approach to the analysis of time series data about reputation of firms. A standard hierarchical clustering method is used and a new measure of distance between time series is proposed. The newly introduced measure is based on qualitative analysis of time series data. The approach is evaluated on a task of clustering Slovenian firms according to the pattern of change of their reputation through years.

1. Introduction

Undoubtedly, reputation of an firm is a dynamical concept. It is usually measured once per year, but most of the analysis methods explore the time local relation of reputation to various financial and performance indicators of the firm. However, time local analysis gives no insight into dynamic change of reputation through years. The temporal change of reputation indicators is usually analyzed with different regression tools (Hammond and J.W. Slocum, 1996).

In this paper, we aim to analyze time series data about the dynamic change of the reputation of firms through years. The presented approach to analysis of reputation data is based on a clustering methodology. Following the clustering methodology, groups of firms with similar patterns of temporal change of their reputation are created. Thus, the key concept in clustering is the measure of distance between firms, or more precisely distance measure between time series that reflects temporal changes of reputation. Most commonly used distance measures for time series analysis are based on the correlation coefficient (Mantegna, 1999; Ormerod and Mounfield, 2000). However, the correlation coefficient is very problematic in cases when we are dealing with very short time series, measured at ten or less time points. To address this problem, we propose the use of an alternative measure of qualitative distance between time series.

We evaluate the performance of the proposed approach on the task of clustering Slovenian firms based on their reputation. The dataset contains data about 113 firms in Slovenia from 1996 to 2000. The data have been collected using the standard Computer Assisted Telephone Interviewing (CATI) method. A quota sample of 760 to 840 Slovenian managers are interviewed about their subjective perception of the reputation of firms. The dataset also includes data about yearly financial performance and advertising investments from the FIPO (fip, 2000) and IBO (ibo, 2000) databases.

The paper is organized as follows. The hierarchical clustering methodology is presented in Section 2. Section 3 introduces the qualitative distance measure between short time series. The results of clustering time series data about the reputation of Slovenian firms are presented and analyzed in Section 4. Section 5 concludes the paper with a summary and directions for further work.

2. Hierarchical clustering

Clustering is an unsupervised learning method. Given data about a set of objects, a clustering algorithm creates groups of objects following two criteria. First, objects are similar (or close) to the other objects from the same group (internal cohesion) and dissimilar (or distant) from objects in the other groups (external isolation).

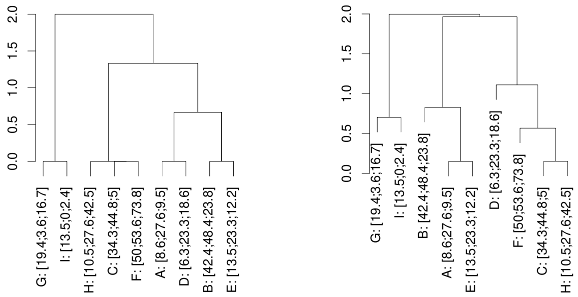
A particular class of clustering methods, studied and widely used in statistical data analysis (Sokal and Sneath, 1963; Hartigan, 1975) are hierarchical clustering methods. The hierarchical clustering algorithm starts with assigning each object to its own cluster, and iteratively joins together the two closest (most similar) clusters together. The distances between objects are provided as input to the clustering algorithm. The iteration continues until all objects are clustered into a single cluster. The output of a hierarchical clustering algorithm is a hierarchical tree or dendrogram.

Two examples of dendrograms obtained by clustering 9 objects are presented in Figure 1. Dendrogram is a binary tree where the initial clusters, consisting of one element only, form the leaves of the tree. Each internal node represents a cluster that is formed by joining together objects from the two clusters corresponding to the children nodes. The height of the node is proportional to the distance between the joined clusters. For example, clusters $\{G\}$ and $\{I\}$ are joined together at height 0 in the dendrogram on the right-hand side of Figure 1, and at height 1 in the dendrogram on the right-hand side of Figure 1.

In the very last step of the clustering, a number of clusters are obtained from the dendrogram. This is done by cutting the dendrogram at a given height. Cutting a single dendrogram at different heights produces different numbers of clusters. For example, cutting the dendrogram on the left-hand side of Figure 1 at height 1.5 produces the following 3 clusters: $\{G, I\}$, $\{H, C, F\}$ and $\{A, D, B, E\}$. Analogously, cutting the dendrogram on the right-hand side of Figure 1 at height 1.4 produces the following 3 clusters: $\{G, I\}$, $\{B, A, E\}$ and $\{D, F, C, H\}$.

The optimal "cut point" that produces clusters with maximal internal cohesiveness and minimal external isolation is where the difference between the height of two successive nodes in the dendrogram is maximal. The dendrogram on the left-hand side of Figure 1 has three equally good cut points (first between 0 and

Figure 1: Two example dendrograms obtained with clustering 9 time series of length 3 using a qualitative distance measure (left-hand side) and a correlation based distance measure (right-hand side).



1, second between 1 and 2 and third between 2 and 3), whereas the one on the right-hand side has only one optimal cut point (any height between between 1.25 and 1.75).

3. Distance measures for short time series

There is a very important question about clustering that has remained unanswered in Section 2.: "how should we measure the distance between objects and between clusters of objects?" Following the fact that in this study we analyze time series data, we are interested in measure of distance between time series.

Most commonly used distance measures are the ones that define distance between two n-dimensional vectors of real numbers. Examples of these are Euclidean and Manhattan distances. These measures are not appropriate for clustering time series, because they mainly capture the difference in the scale and baseline (order of magnitude difference) between objects. Instead of that, in clustering time series, we are more interested in the shape of the time change of the value through time.

A better alternative is to use a correlation based distance measure. The correlation coefficient $r(X, Y)$ between two time series X and Y, calculated as

$$r(X,Y) = \frac{E[(X - E[X]) \cdot (Y - E[Y])]}{E[(X - E[X])^2] \cdot E[(Y - E[Y])^2]},$$

where $E(V)$ is used to denote the expectation (i.e., mean value) of V, measures the degree of linear dependence between X and Y. Values of $r(X, Y)$ close to zero denote that there is a low degree of linear dependence between X and Y. On the other hand, values close to ± 1 denote a high degree of linear dependence. In terms of shapes of the X and Y, the value of $r(X, Y)$ has the following intuitive meaning. Values close to -1 means that X and Y have "mirrored" shapes, r close to 0 means that shapes are unrelated (and consequently dissimilar) and r close to 1 means that the shapes are very similar. Following this intuitive interpretation of correlation we can define the following distance measure between time series (Ormerod and Mounfield, 2000):

$$D_r(X,Y) = \sqrt{2 \cdot (1 - r(X,Y))}.$$

However, this correlation based distance measure has two drawbacks. First, it is well known that the correlation coefficient is very poorly estimated when we have small number of observations (i.e., short time series). Second, it is capable of capturing only the linear aspect of dependence or relation between time series. Two time series that are non-linearly related to each other will be distant from each other, regardless of the similarity of their dynamic change through time.

The distance measure that we propose here is based on a qualitative analysis and comparison of the shape of the time-series. Consider the two time series X (left-hand) and Y (right-hand) of length five, presented in Figure 2. We choose a pair of time points i and j and we observe the qualitative change of the value of X and Y. Three possible values of qualitative change $q(X_i, X_j)$ (as well as $q(Y_i, Y_j)$) can be distinguished: increase, when $X_i > X_j$; no-change when $X_i = X_j$ and decrease $X_i < X_j$[1]. For example, consider the change between the first and third time point in X and Y from Figure 2: they both increase. On the other hand, if we consider the change between the second and third point, we observe that while X decreases, Y increases. Now, we can calculate the qualitative difference between X and Y by summing up the differences for all the pairs of time points:

Table 1: Definition of the `Diff` function.

	$\mathtt{Diff}(q_1, q_2)$	q_1: increase	no-change	decrease
q_2	increase	0	0.5	1
	no-change	0.5	0	0.5
	decrease	1	0.5	0

$$D_q(X,Y) = \frac{4}{N \cdot (N-1)} \cdot \sum_{i<j} \mathtt{Diff}(q(X_i, X_j), q(Y_i, Y_j)),$$

where $q(V_i, V_j)$ is used to denote the qualitative change of V and `Diff` is a simple function that defines the difference between three possible values of qualitative change. The factor $\frac{4}{N \cdot (N-1)}$ is used to normalize the values of the distance measure in the range $[0, 2]$ which equals the range of values of the correlation based distance D_r.

The `Diff` function is defined in Table 1. The definition simply specifies that the difference between increase and decrease values equals 1, whereas the difference between increase (or decrease) and no-change values equals 0.5.

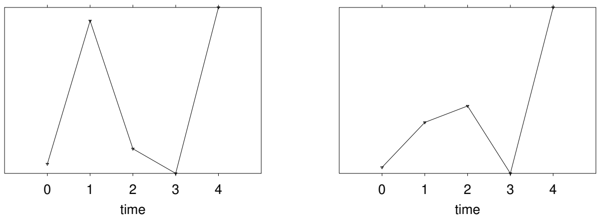
Roughly speaking, D_q counts the number of disagreements of change of X and Y. It equals 0 if both time series increase and decrease at same time. In the qualitative reasoning methodology QSIM (Kuipers, 1986) this is denoted by using the qualitative constraint $M^+(X, Y)$. The maximal distance of 2 is obtained in cases when X decreases wherever Y increases. The QSIM notation for this situation is $M^-(X, Y)$.

The proposed qualitative distance measure does not have the drawbacks of the correlation based measure, mentioned above. First, it can be calculated on very short time series, without decreasing the quality of the estimate. On the other hand, calculating D_q for pairs of very long time series can be impractical for time complexity reasons.[2] Second, it captures the similarity between patterns of change of the time series, regardless of whether

[1]Note the strict definition of the no-change situation is used here for simplicity. In reality, we use a threshold ($|X_i - X_j| < \epsilon$) to test the equality of X_i and X_j. Alternatively, if more then one measurement for X_i is available, a statistical test of the significance of change can be applied for the same purpose.

[2]Note that the time complexity of calculation of D_q is quadratic in the length of the time series, due to the fact that all possible pairs of time points are considered.

Figure 2: Two example time series of length 5. The correlation based distance between them equals 0.694, whereas the qualitative distance equals 0.2.



the nature of the dependence between them is linear or non-linear. An illustration of the advantage of using the qualitative distance measure is given in Figure 2. Although the pattern on the left-hand side is very similar to the pattern on the right-hand side, the distance between them is rather high according to the correlation based measure (0.694). It is three times higher than the one measured by the qualitative distance measure (0.231).

4. Clustering of Firms Reputation Data

We applied the presented approach to the task of clustering firms based on their reputation. In this section, we first describe the dataset and the methodology used for measuring the reputation and collecting other data about Slovenian firms. We then present the results of clustering and analyze the obtained clusters.

4.1. Experimental Methodology

The dataset contains data about 113 firms in Slovenia covering the period 1996 to 2000. For each firm, the dataset includes measurements of three groups of descriptors. The first group includes time-invariant general data such as the name of the firm and (industrial or service) area of the firm activity. The second group of features include estimates of the recognition and reputation of the firms. These estimates are based on the customers' subjective perception of the reputation. The data about this group of features have been collected using the standard Computer Assisted Telephone Interviewing (CATI) method. Further details about the CATI methodology are given below. Finally, the third group of features includes yearly financial performance indicators, such as the firm's equity, revenues, net profit and number of employees, as well as advertising investments data. The data about this group of features have been collected from the publicly available databases FIPO (fip, 2000) and IBO (ibo, 2000). The features in this group are measured once per year for the period 1996 to 1999.

Since 1995, Kline and Kline marketing agency has conducted annual surveys that assess corporate reputation of 255 largest companies in Slovenia. The standard CATI methodology is used to colect answers from quota sample of 760 to 840 Slovenian managers. The corresponding data are stored in the computer program QA that was designed to facilitate the questioning process. After the data gathering process is completed, the program exports the collected data in a standardized format, so that it can be imported to other statistical or data mining software for further processing.

One of the most important features of the QA program is to lower the complexity of estimating the image of firms. Note that the idea of asking each individual to evaluate all of the firms would surely degrade his or her initiative to cooperate. Therefore, a careful experimental design is required before one can start questioning people. From our experience, as well as from theoretical psychology, we concluded that in order to obtain valuable information, each individual could subjectively evaluate at most 15 firms. As s result, QA program in each run randomly selects a permutation of 15 firms, balancing the overall frequency rate for each firm.

We performed two clustering experiments. The first is on a dataset for the period of five years from 1996 to 2000 that contains complete data for 82 firms. Since many fields describing the years 1996 and 1997 were missing, we repeated a simplified version of the first experiment on a dataset for the period of three years from 1998 to 2000 that contains complete data for 106 firms. In both experiments, hierarchical clustering with the qualitative distance measure was used. For the purpose of comparison of results, we also used a correlation based distance measure.

4.2. Experimental Results

In both experiments, the presented approach generated 4 clusters of firms. The summary of the clusters for both experiments is presented in Table 2.

The first two rows in each table are the sizes of clusters (the number of firms in each cluster) and the distribution of firms among clusters. Further down the rows in both tables, the average values of reputation and some financial and performance indicators are presented.

4.3. Preliminary Analysis of the Results

The first observation is that a small number of clusters is obtained in both cases. This compares favorably with the number of clusters generated with the correlation based distance measure. The latter generated 16 clusters for the first dataset and 8 clusters for the second dataset.

The four clusters obtained are immediately recognizable in both cases. Of course, it is easier to recognize the patterns for the second dataset with shorter time series. In the first dataset, the patterns become more complex, but still they are compact enough to still be recognizable. In both cases, we can identify the following four important and relevant groups of firms. The first group, metaphorically named gazelles, includes firms with rapidly increasing growth of reputation. The firms in the second group of mugwumps have stable, but slower growth of reputation that is on average higher then the reputation of firms in the fist group. The third group of decays includes mostly firms with constant decrease of the reputation. The fourth group includes lingerers, i.e., firms with constant ups and downs. Of course, in each of the described groups there are exceptions from the general pattern. The analysis of exceptions is more complex and requires a separate analysis of data for each exceptional firm.

We also analyzed the distribution of values of other financial and performance indicators among clusters. From the table, we observe that the firms in the fourth cluster obtained in the first dataset (the one of the mugwumps) are highly reputable (with average reputation of 34.32 versus the overall reputation of 28.53), but also have higher average equity (14M vs. 9M), assets (22M vs. 5M) and equity on assets (66.68 vs. 60.04). It is interesting that the firms in this cluster are not the ones with the highest average advertising budget. The high advertising budget is typical for the firms in the second cluster of gazelles.

Please note that the observations presented here are preliminary. For more conclusive statements, a test of statistical signifi-

Table 2: Summary of the clusters generated on both experimental datasets.

first dataset	cluster 1	cluster 2	cluster 3	cluster 4	full dataset
size	10	29	34	33	106
distribution	9.43%	27.36%	32.07%	31.14%	100%
reputation	15.32	31.67	24.11	34.32	28.53
employees	690.12	657.33	912.08	1,192.05	911.78
equity	3,221,372.33	6,923,234.12	8,310,799.32	14,835,049.93	9,686,419.34
assets	8,728,310.46	11,743,432.20	13,426,307.40	22,726,639.07	15,727,797.17
equity on assets	42.68	56.24	61.96	66.68	60.04
revenues	9,758,372.67	13,750,380.12	12,131,758.02	25,038,866.90	16,877,261.45
net profit	1,683,917.50	378,695.03	436,917.19	1,049,282.17	759,206.09
net profit on rev.	145.88	3.89	4.46	5.86	17.32
advertising	8,329,958.17	107,488,647.38	31,849,193.61	92,239,627.24	69,261,171.31

second dataset	cluster 1	cluster 2	cluster 3	cluster 4	full dataset
size	13	30	25	14	82
distribution	15.85%	36.58%	30.49%	17.08%	100%
reputation	25.72	33.94	36.62	23.68	31.70
employees	747.64	1,081.46	972.47	1,320.76	1,032.58
equity	5,926,813.87	12,507,134.41	12,387,730.71	11,200,748.06	11,202,319.84
assets	11,828,923.13	20,833,307.33	16,633,921.88	18,131,153.82	17,799,067.92
equity on assets	53.31	58.67	73.87	64.04	62.86
revenues	10,466,048.83	21,506,881.25	23,223,478.82	14,270,215.67	18,954,032.25
net profit	575,692.47	881,447.59	819,185.20	470,787.58	744,411.52
net profit in rev.	4.20	4.99	6.78	3.52	5.09
advertising	80,785,072.90	121,001,002.85	71,855,421.89	32,405,566.90	84,672,149.32

cance of differences should be performed.

5. Conclusions and Further Work

In this paper, we present a clustering approach to analysis of time series data about the dynamic change of the reputation of firms through years. We use hierarchical clustering to obtain groups of firms with similar patterns of change of reputation through years. We proposed a novel measure of distance between time series that is especially suitable for very short time series, where the use of commonly used correlation based distance measure is not appropriate.

The preliminary analysis of the obtained clusters shows the usability of the approach. The number of clusters is small and they reflect representative groups of Slovenian firms. This is in contrast to clustering with a correlation based distance measure where many clusters were generated.

However, the analysis of the discussion of the results presented here is preliminary. Many aspects of the analysis should be improved. First, we observe considerable amount of variance of the values of different firms' features among clusters. The statistical significance of this variance should be tested. Also, the reasons for this variance should be further explored and explained.

Furthermore, the approach presented here can be a first step towards more extensive analysis of dynamic change of the reputation of firms. One line of extensions would be towards practical applications by building predictive models of reputation, that can be widely used by managers and analysts for prediction and planing. Another line of further work is to analyze the "buffer" hypothesis about the delay between reputation change and the consequential, delayed, change of other financial and performance indicators of the firm. For this purpose, the delay operator between time series should be taken into account in the further development of the methodology.

Finally, clustering methodology has already been used for analysis of reputation of German Firms in (Dunbar and Schwalbach, 2000). They use alternative approach, where each year firms are clustered according to their level of reputation. Then the dynamic change of cluster membership can be observed in order to analyze the dynamic change of reputation through years. One of the goal of further work is to compare this approach to the clustering time series approach presented in this paper.

Acknowledgments

The work reported in this paper was supported by the Slovenian Ministry of Education, Science and Sport, and the IST-1999-11495 project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise.

6. References

R. L. M. Dunbar and J. Schwalbach. 2000. Corporate reputation and performance in Germany. *Corporate Reputation Review*, 3(2):115–123.

2000. FIPO Database. Gospodarski Vestnik, Ljubljna, Slovenia. `http://www.gvin.com/FIPO/`.

S. A. Hammond and Jr. J.W. Slocum. 1996. The impact of prior firm financial performance on subsequent corporate reputatation. *Journal of Business Ethics*, 15:159–165.

J. A. Hartigan. 1975. *Clustering Algorithms*. Wiley, New York.

2000. IBO Database. Media Research Institute Mediana, Ljubljna, Slovenia. `http://www.mediana-irm.si/eng/02_02.html`.

B. Kuipers. 1986. Qualitative simulation. *Artificial Intelligence*, 29(3):289–338.

R. N. Mantegna. 1999. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–198.

P. Ormerod and C. Mounfield. 2000. Localised structure in the temporal evolution of asset prices. In *New Approaches in Financial Economic Conference*, Santa Fe, NM. Available for download from `http://www.quantnotes.com/publications/volterra/sf21092000.pdf`.

R. R. Sokal and P. H. A. Sneath. 1963. *Principles of Numerical Taxonomy*. Freeman, San Francisco.