Computer Networks 57 (2013) 3025–3038

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Data clustering based on correlation analysis applied to highly variable domains

Stefania Tosi *, Sara Casolari, Michele Colajanni

Department of Information Engineering, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Available online 14 July 2013

Keywords:
Traffic clustering
High variability
Correlation index
Network management

ABSTRACT

Clustering of traffic data based on correlation analysis is an important element of several network management objectives including traffic shaping and quality of service control. Existing correlation-based clustering algorithms are affected by poor results when applied to highly variable time series characterizing most network traffic data. This paper proposes a new similarity measure for computing clusters of highly variable data on the basis of their correlation. Experimental evaluations on several synthetic and real datasets show the accuracy and robustness of the proposed solution that improves existing clustering methods based on statistical correlations.

1. Introduction

Clustering is a widely adopted approach for augmenting the level of knowledge on rough data. The goals of clustering applied to computer and network datasets can be different, going from Web sites characterization [1], classification of users navigation patterns [4], network traffic classification and management [2]. For example, many network management goals such as flow prioritization, traffic shaping and policing, and diagnostic monitoring as well as many network engineering problems, such as workload characterization and modeling, capacity planning, and route provisioning may benefit from traffic clustering [2].

In this paper, we are interested in correlation-based clustering algorithms applied to highly variable time series. This set of algorithms (e.g., Pearson product moment [7], Spearman and Kendall ranks [8,9]) consider that time series are similar if they exhibit some degree of statistical inter-dependency, and differ from other popular approaches using some geometrical distance (e.g., Euclidean distance [6], cosine distance [5]) as their *similarity measure*. The reason of focusing on correlation similarity measures is distance functions are not always adequate in capturing dependencies among the data. In fact, strong dependencies may exist between time series even if their data samples are far apart from each other as measured by distance functions [3]. In the next section, we will support this statement through a network related example.

The choice and the performance of the similarity measure impact the quality of any clustering algorithm. The better the accuracy and robustness of the measure in finding similarity, the better the quality of the clustering model. Existing correlation indexes are accurate and robust in disclosing similarity except when time series exhibit high variability. This is the case of most traffic data that are highly variable in terms of number of connections, request inter-arrivals, flow sizes (e.g., [13,16,14]). In these scenarios, popular correlation indexes, such as the Pearson coefficient [7], the Spearman rank [8], the Kendall rank [9], and the Local Correlation index [10], show poor results because they are unable to capture correlations even when they exist.

We propose a new similarity measure that is able to disclose correlation even when time series are characterized by high variability. The accuracy and robustness of the proposed correlation index is achieved through an

* Corresponding author. Address: Via Vignolese 905/B, 41125 Modena, Italy, Tel.: +39 0592056273; fax: +39 0592056129.

E-mail addresses: stefania.tosi@unimore.it (S. Tosi), sara.casolari@unimore.it (S. Casolari), michele.colajanni@unimore.it (M. Colajanni).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2013.07.004

original approach that separates trend from perturbation patterns, and evaluates correlation by computing the similarity of trend patterns. On this basis, clustering models can group time series presenting similarity also when characterized by high variability, such as network traffic [16], workloads [15], and data center resource metrics [11]. Such data may have strong correlations that are masked by perturbations. When some correlations exist, the similarity measure we propose is able to identify them and clustering algorithms can group time series accordingly. The improvements with respect to the state of the art are shown on synthetic and real datasets characterized by high variability.

The remainder of this paper is organized as follows. Section 2 defines the problem of correlation clustering for highly variable datasets. Section 3 presents the proposed algorithm. Section 4 compares the performance of different correlation indexes applied to synthetic time series that represent a fully controlled scenario for evaluation. Section 5 evaluates the proposed algorithm on real scenarios. Section 6 concludes the paper with some final remarks.

2. Problem definition

We define the clustering process based on similarity by considering a dataset $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. For example, it contains all time series of a monitored network, where each time series $\mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]$ is a vector containing a time-ordered discrete sequence of traffic data sampled once. We are interested in partitioning the N time series into K clusters $\mathcal{C} \equiv \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ $(K \leqslant N)$, such that:

1. $\mathcal{C}_i \neq \emptyset, i = 1, \ldots, K$;
2. $\bigcup_{i=1}^{K} \mathcal{C}_i = \boldsymbol{X}$;
3. $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset,\ i,j = 1, \ldots, K$ and $i \neq j$.

The clustering algorithm requires the choice of a similarity measure determining groups of time series so that the similarity between time series within a cluster is larger than the similarity between time series belonging to different clusters. As a similarity measure, we adopt the *correlation index* ρ between two time series $\mathbf{x}_i$ and $\mathbf{x}_j \in \boldsymbol{X}$, where the absolute value of ρ ranges between 0 and 1. When $\rho = 0$, there is no correlation between the two time series, while $\rho = 1$ indicates a complete correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. The literature offers several guidelines for the best interpretation of the value of the correlation measure [17,7], but all criteria depend on the context and purposes of the analysis. In this paper, we do not refer to a specific traffic scenario, hence we can adopt the most general interpretation indicating a *strong correlation* when $\rho > 0.5$, and a *weak correlation* for $\rho \leqslant 0.5$ (e.g., [17]). Different choices for the threshold do not impact the main conclusions of this paper.

This paper proposes a new similarity measure that is able to determine correlation clustering even in datasets exhibiting a high degree of variability, where existing correlation indexes (e.g., [7,10,8,9]) are not accurate. High variability is a typical phenomenon in network-related time series [16] in which most observations take values around the time series trend (*trend pattern*) and some observations depart from it with appreciable frequency, even by assuming extremely large values with non-negligible probability (*perturbation pattern*). Trend patterns represent the tendency of a time series that may be related to the other time series, while perturbation patterns consist of random observations hiding trends. In this paper, we use the standard deviation as the measure of time series variability because a high standard deviation is the most typical trademark of highly variable network measurements [16]. For our purposes, this feature causes trend patterns that are hard to identify because masked by perturbations.

Fig. 1 illustrates some examples of highly variable time series derived from network monitors measuring the number of active connections, active clients and transferred bytes during a day period. In each time series, we can see the presence of different trend patterns during working hours and during the night. These patterns are masked by perturbation patterns. However, there is an evident dependency between the number of active connections and the amount of transferred bytes. In fact, when the number of active connections increases (decreases), the amount of transferred bytes increases (decreases) as well. Despite the variability, we want that a good similarity measure can detect this dependency so to group the two time series in the same cluster. This goal cannot be achieved through distance-based similarity measures because the distance between the sample values of the two time series is not always close, hence the two time series cannot be clustered together through a traditional distance-based clustering model. For this reason, we prefer to consider correlation as the similarity measure, and we propose a correlation-based clustering model that is able to disclose dependency even in highly variable scenarios where distance-based clustering models do not work.

The ability of a correlation index in detecting similarity among correlated time series is measured in terms of *accuracy*. The ability in guaranteeing a stable correlation index when conditions do not change is measured in terms of *robustness*. In the case of highly variable time series, the most popular correlation indexes are affected by two main problems:

1. low accuracy, since they are unable to detect similarities even among correlated time series;
2. low robustness, since they do not guarantee a stable evaluation of the correlation index, even when the relationships between the time series do not change.

Let us give an example of the above problems by referring to the time series shown in Fig. 1 that reports the number of active connections and transferred bytes monitored during a 12-h period in Fig. 2(a). We evaluate the correlation index between the two time series through the Pearson correlation index [7], and we report the results in Fig. 2(b). (The Pearson index is used as an example, but other existing models do not change the conclusions.) Despite the relationship between the two time series, the Pearson model is affected by both issues reported earlier: its results are characterized by low accuracy because the Pearson correlation index remains lower than 0.5 during