A Dimensionality Reduction Technique for Efficient Similarity Analysis of Time Series Databases

Vasileios Megalooikonomou Guo Li Qiang Wang
CIS Department, Temple University
1805 North Broad Street, Philadelphia, PA 19122
1.215.204.5774
{vasilis, gli00001, wang32}@temple.edu

ABSTRACT
Efficiently searching for similarities among time series and discovering interesting patterns is an important and non-trivial problem with applications in many domains. The high dimensionality of the data makes the analysis very challenging. To solve this problem, many dimensionality reduction methods have been proposed. PCA (Piecewise Constant Approximation) and its variant have been shown efficient in time series indexing and similarity retrieval. However, in certain applications, too many false alarms introduced by the approximation may reduce the overall performance dramatically. In this paper, we introduce a new piecewise dimensionality reduction technique that is based on Vector Quantization. The new technique, PVQA (Piecewise Vector Quantized Approximation), partitions each sequence into equi-length segments and uses vector quantization to represent each segment by the closest (based on a distance metric) codeword from a codebook of key-sequences. The efficiency of calculations is improved due to the significantly lower dimensionality of the new representation. We demonstrate the utility and efficiency of the proposed technique on real and simulated datasets. By exploiting prior knowledge about the data, the proposed technique generally outperforms PCA and its variants in similarity searches.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *data mining;* H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval – *clustering, search process;*

General Terms
Algorithms, Management, Performance, Experimentation.

Keywords
Time series, dimensionality reduction, data mining.

1. INTRODUCTION
The problem of retrieving similar time sequences may be stated as follows: Given a query q, a database S: $S_1,S_2,...,S_N$, a distance measure D and a threshold ε, find the sequences R in S that are within distance ε from q. More precisely, $R = \{S_i \in S | D(q,S_i) \leq \varepsilon\}$. In a variant of this problem, no threshold is given, instead the closest neighbors of the query series are to be found. To compare two given time series, a suitable measure of similarity should be given. The Euclidean distance is most often used.

In many situations, the high dimensionality of time series makes the distance calculation very inefficient. Promising techniques include those based on dimensionality reduction and multidimensional indexing. An efficient approach is based on piecewise constant approximation (PCA) or piecewise aggregate approximation (PAA). Yi and Faloutsos [7] and Keogh et al [2] proposed to divide each sequence into k segments of equal length and to use the average value of each segment as a coordinate of a k-dimensional feature vector. Recently, a symbolic PAA was also introduced [3].

In this paper, we introduce a new method to efficiently reduce the dimensionality of time series. Our work is motivated by the observation that the mean value that is being used to approximate each equi-length segment in PCA and PAA is the best one can do for a piecewise approximation if there is no prior knowledge about the data or a method needs to be independent of the data. The method proposed here is also based on segmentation of a sequence but extends PCA by allowing a more flexible approximation of each segment, using ideas from data compression and in particular the vector quantization technique, effectively representing a long time series with a symbolic representation of much lower dimensionality. In addition to being comparable to the other popular methods in terms of complexity, the proposed approach demonstrates advantages of closer approximation of original time series and higher accuracy in time series matching.

CIKM'04, November 8-13, 2004, Washington, DC, USA.
ACM 1-58113-874-1/04/0011.

2. METHODOLOGY
The proposed approach, Piecewise Vector Quantized Approximation (PVQA), partitions a sequence into equi-length segments and uses vector quantization (VQ) to represent each segment with the closest codeword from a codebook. VQ is widely used in signal compression and coding; it is a lossy compression method based on the principle of block coding [1].

During a training phase, a codebook $C=\{c_1,c_2, ..., c_s\}$ of size an arbitrary integer s ($s \geq 2$) is created. A time series $X = x_1, x_2, ..., x_n$ of length n is represented with a vector $X'=x'_1,x'_2,...,x'_w$ of length w ($w << n$) by being segmented into w equal size segments. The i-th element of X' is:

$$x'_i = \arg\min_k (\mathrm{D}(\mathrm{SEG}_i, \mathrm{c}_k);\ \mathrm{k} = 1, \ldots, s)$$

where SEG_i is the i-th segment in X, D is the distance measure (e.g., Euclidean distance), and c_k is the k-th codeword in C.

2.1 Codebook Generation
Each time series in a training set T is partitioned into a number of segments of a fixed length l and each segment forms a sample that is used to generate the codebook. In order to get the key-sequences (codewords) and build the codebook we apply the Generalized Lloyd Algorithm (GLA) [4].

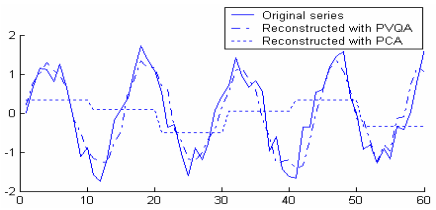

Figure 1. A time series and its reconstructions

2.2 Data Encoding

In the process of encoding, every series is decomposed into subsequences of length l (same as the length of the codewords). For each subsequence the closest entry c_k in the codebook is found and its index k is stored. So, the new representation of a time series is a vector of indices to codewords. Figure 1 shows an original time series and its reconstruction given a certain codebook by concatenating the corresponding codewords. For comparison, we also show the reconstruction using PCA. PVQA has more flexibility than PCA to approximate the original time series arbitrarily close through the adjustment of not only the number of segments, but also the size of the codebook.

2.3 Distance Measures

Using PVQA, a time series is represented as $X'= x'_1,x'_2,...x'_w$, and correspondingly a query is represented as $Q'=q'_1,q'_2,...q'_w$. Each x'_i and q'_i ($1\le i \le w$) is an index corresponding to a codeword in the codebook. Since the approximate representation for a time series X (Q) is the concatenation of all the codewords corresponding to x'_i (q'_i), we can sum up the distance between each pair of x'_i and q'_i and get a rough distance between the two time series:

$$\text{RoughDist}(\ X, Q) = \sqrt{\sum_{i=1}^{w} (D(x_i^{'}, q_i^{'}))^2}$$

The distance between each pair of codewords can be pre-calculated and stored. The space complexity of the distance matrix is $O(s^2)$ and the time complexity of computing the rough distance between two time series is $O(w)$.

3. EXPERIMENTS

To evaluate the proposed method, we performed experiments in best match searching, i.e., given a query sequence, find the best k matches in a database. The evaluation metric we used was the percentage of the results that fall in the same class as the query. We compared the efficiency and accuracy of our method to that of two other piecewise dimensionality reduction techniques: PCA and symbolic PAA. For fairness, we used the same reduced dimensionality for all methods. The accuracy of the Euclidean distance on the original time series (Naïve approach) was also calculated. Using the *tightness of approximation* defined as RoughDist(X,Q) / D(X,Q), we ran experiments on several synthetic and real datasets and chose w=6 and s=16 as a nice tradeoff between accuracy and efficiency. In order to assure that the experimental results are reliable, we applied 5-fold cross-validation and the datasets were preprocessed with Z-normalization.

In Figure 2 we show the experimental results on a synthetic dataset, SYNDATA [6], and on a real dataset, GENE [5]. SYNDATA contains 600 examples from 6 classes of control charts (each having 60 points). For this dataset, we used k = 2, 5, 8, 10, 15, 20. GENE is a subset of the NCI60 gene expression data from the National Cancer Institute. It consists of 36 cancer cell lines of 6

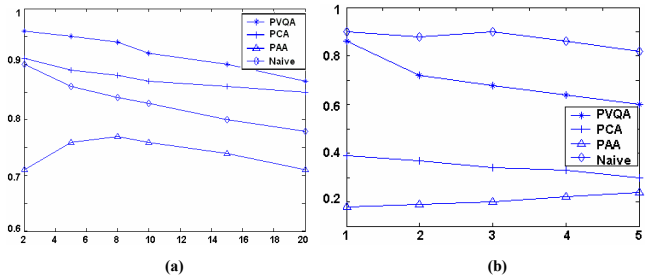

Figure 2. Matching results on SYNDATA (a) and GENE (b)

different kinds of cancer and each series has the expression values of 1375 genes. For GENE, we used k= 1, 2, 3, 4, 5.

As shown in Figure 2, the matching accuracy of PVQA is close or even better than that of Euclidean distance and is much better than the results obtained with PCA or PAA. With PVQA, while the outline of the original time series is kept, noise that may affect the calculation of similarities between different time series is removed and this leads to the improved accuracy.

4. CONCLUSIONS

We have proposed a novel symbolic representation of time series that effectively reduces the dimensionality improving the efficiency of calculations in similarity searches. The proposed PVQA approach is a natural extension of the piecewise constant approximation schemes proposed earlier. By exploiting prior knowledge about the data and allowing the use of a very tight approximation of the Euclidean distance we were able to improve performance in time series similarity analysis over previously proposed methods. Moreover, the proposed representation is symbolic and potentially allows the application of text-based retrieval techniques into the similarity analysis of time series.

5. ACKNOWLEDGMENTS

This material is based upon work supported by NSF under Grant No. IIS-0237921, by NIH under Grant No. R01MH68066-01A1 (funded by NIMH, NINDS, and NIA) and by the Pennsylvania Department of Health.

6. REFERENCES

[1] Gersho, A. & Gray R. M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic, Boston.

[2] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. (2000). "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases", *Knowledge and Information Systems* 3(3): 263-286.

[3] Lin, J., Keogh, E., Patel, P. & Lonardi, S. (2002). "Finding motifs in time series", 2nd Workshop on Temporal Data Mining at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23 - 26. Edmonton, Alberta, Canada.

[4] Lloyd, S. P. (1982). "Least squares quantization in PCM", *IEEE Transactions on Information Theory,* IT(28), pp. 127-135.

[5] Stanford Genomic Resources. http://genome-www.stanford.edu/nci60

[6] UCI KDD Archive. http://kdd.ics.uci.edu

[7] Yi, B-K & Faloutsos, C. (2000). "Fast Time Sequence Indexing for Arbitrary Lp Norms", in Proceedings of the VLDB, Cairo, Egypt, pp. 385 - 394.