Clustering Financial Time Series on CUDA

Gábor Bényász and László Cser

Clustering in financial time-series databases has received significant attention lately. Differently from the normal clustering algorithm the main challenge of the time-series clustering is the high dimensionality. The fact, which helped to overcome this awkwardness, was the development of the time series models which enabled the utilization of clustering of time series by compressing the dimensionality of time series into parameterised expression which can be compared and clustered.

The first models based upon moving average and autoregression of the time series called AR, MA and ARIMA [2] models. There were introduced some clustering model [1], [4] those ARIMA time-series which are intend to use the Linear Predictive Coding (LPC) cepstrum of time series as Euclidean distance between the LPC cepstrals of the time series. More developed model of time-series, called GARCH [3], models take also account the autoregressive conditional heteroscedasticity of the time series and therefore is able to interpret nonstationarity time-series. As clustering of GARCH models of time series has not been investigated so far the next logical step could have been to cluster the GARCH based time-series, even DWT (Discrete Wavelet Transformation) was chosen for modelling time-series which can be clustered after that. We chosed that because of the shortcomings of the GARCH models as it cannot handle adequately and entirely the shock-like movements of the time-series (which is essential character of the financial data) which followa long-tail distribution. Similarly DFT (Discrete Furier Transformation) has been also heavily investigated [6] [7] for time series clustering, but [8] showed that DWT is significantly faster to model and provide a multi-resolution decomposition.

DWT has been intensively used in technical and natural sciences for decades, but using it for financial purposes is a very new initiative. Using DWT time series are projected into the time-frequency plane of sliding windows. The Wavelet coefficients hold the compressed characterization of the time series and form the dimension of time-series clustering. The major differences of using the DWT for comparing the time-series are rooted in the usage of the coefficients (first k, last k, largest k, adaptive [9] coefficients) for further reducing the dimensionality because in high dimensional spaces the distance between the nearest and the farthest neighbour gets increasingly smaller [5], making it impossible and meaninglessly to cluster. We consider the use the adaptive method discussed in [9] to reduce the number of the coefficients. As a side effect the high dimensionality of the clustering algorithm can slow down so that even becomes useless. To overcome this performance related difficulties CUDA was intend to use as clustering algorithms could be massively parallelised [10, 11, 12, 13]. Not knowing the initials numbers of the clusters we considered using iterative k-Means algorithm. To enhance the implementation the wavelet transformation was also carried out via CUDA as Wavelet algorithms are naturally parallel. For example, if enough processing elements exist, the wavelet transform for a particular spectrum can be calculated in one step by assigning a processor for every two points. The proposed algotrithm are attempted to run and evaluate using of closing prices of stocks listed on the NYSE. The method for the determination of the sliding window is still subject of the further investigation.

References

[1] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta: Distance Measures for Effective Clustering of ARIMA Time-Series. *In the Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01), San Jose, CA, November 29-December 2, 2001, pp. 273-280.*

[2] Brockwell, Peter J. and Davis, Richard A.: *Time Series: Theory and Methods, 1987, Springer-Verlag.*

[3] Engle R. F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom ination. *Econometrica 50, 1987, 987-1007.*

[4] Bagnall A. J., Janacek G. J.: Clustering time series from ARMA models with clipped data. *In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Seattle, WA, USA, August 22 - 25, 2004). KDD '04. ACM, New York, NY, 49-58.*

[5] Beyer K., Goldstein J, Ramakrishnan K, Uri S.: When is nearest neighbor meaningful? *Lecture Notes in Computer Science, 1540:217-235, 1999.*

[6] Agrawal R, Faloutsos C., and Swami A. N.: Effcient Similarity Search In Sequence Databases. *In D. Lomet, editor, Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO), pages 69-84, Chicago, Illinois, 1993. Springer Verlag.*

[7] Rafiei D. and Mendelzon O. A.: Similarity-based queries for time series data. *In Proceedings of the 1997 ACM SIGMOD international Conference on Management of Data (Tucson, Arizona, United States, May 11 - 15, 1997). J. M. Peckman, S. Ram, and M. Franklin, Eds. SIGMOD '97. ACM, New York, NY, 13-25.*

[8] Wu Y, Agrawal D.,Abbadi A. E.: A comparison of DFT and DWT based similarity search in time-series databases. *In CIKM, pages 488-495, 2000.*

[9] Morche F.: Time series feature extraction for data mining using DWT and DFT, *whitepaper, 2003*

[10] YING, L.: High performance computing with CUDA, *Conference of High Performance Computing, XVII, 2009)*

[11] Zechner M, Granitzer M.: Accelerating K-Means on the Graphics Processor via CUDA, *pp.7-15, 2009 First International Conference on Intensive Applications and Services, 2009*

[12] Shuai, C. et al: Performance Study of General Purpose Appliaction on GPU using Cuda, *Journal of Distributed and Paralell computing, XVI, 2009.*

[13] Shalom, S. A., Dash, M., and Tue, M.: Efficient K-Means Clustering Using Accelerated Graphics Processors. *In Proceedings of the 10th international Conference on Data Warehousing and Knowledge Discovery (Turin, Italy, September 02 - 05, 2008). I. Song, J. Eder, and T. M. Nguyen, Eds. Lecture Notes In Computer Science, vol. 5182. Springer-Verlag, Berlin, Heidelberg, 166-175*