SAX: An innovative outlook to time series representations

Aayush Kumar, Sejal Mohata, Arjun Bahuguna, and Soumyaranjan Sabat

***Abstract*— Predictive analysis on Time Series Data Sets using SAX: The potential to do predictive machine learning on the data generated by connected sensors (temperature sensor, LIDAR etc..) is a feature that is driving the spread of the Internet of Things. Current research in indexing and mining time series data has produced many interesting algorithms and representations.However, it has not led to algorithms that can scale to the increasingly massive datasets encountered in science, engineering, and business domainsPredictive analysis on the time series data can be used to anticipate adverse events, enable early warning systems, improve results, reduce costs and enhance efficiency. Probability is not defined for the time series, Probability can be defined for Symbols. Mapping a time series to a symbol may allow us to assign a probability to the time series sub-sequence. This involves mapping the time series sub-sequence to a symbol in some symbol space. Key advantages of using SAX is that it yields an explainable model. Wherein the result of the model should not be underestimated.**

I. INTRODUCTION

SAX is the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure. In classic data mining tasks such as clustering, classification, summarization, indexing, dimensionality reduction, numerosity reduction etc., SAX is as good as well-known representations such as Discrete Wavelet Transform and Discrete Fourier Transform , while requiring less storage space. In addition, the representation provides solutions to many challenges associated with current data mining tasks. There is great potential for extending and applying the discrete representation on a wide variety of data mining tasks.

A time series is a sequence of pairs where each pair consists of a time index and value. The time index may be implied if there is a constant difference between values. The time series can be segmented into Windows which represent the time series between two-time indices.

The symbol can represent Windows Because symbols in a finite symbol space and have a probability. Symbols are easy to store and manipulate - each symbol can be represented by an integer or a float variable. Creating a few approximation techniques helps to fit the data in main memory.

Advantages of symbolic representation are:

1. Lower bounding of Euclidean Distance
2. Dimensionality Reduction
3. Numerosity Reduction

This representation is known as SAX Symbolic Aggregate Approximation

II. SYMBOLIC AGGREGATE APPROXIMATE

A. About SAX

SAX allows a time series of arbitrary length n to be reduced to a string of arbitrary length w, (w less than n, typically w lesser than n). The alphabet size is also an arbitrary integer a, where a greater than 2.

B. Discretization Process

Our discretization procedure is unique in that it uses an intermediate representation between the raw time series and the symbolic strings. We first transform the data into the Piecewise Aggregate Approximation (PAA) representation and then symbolize the PAA representation into a discrete string. There are two important advantages to doing this:

a) Dimensionality Reduction : We can use the defined and documented dimensionality reduction ability of PAA, and the reduction is automatically carried over to the symbolic representation.

b) Lower Bounding : Proving that a distance measure between two symbolic strings lower bounds the true distance between the original time series is non-trivial. The key observation that Discrete Fourier Transform, Piecewise Linear Approximation, Haar Wavelet Adaptive Piecewise, Constant Approximation allowed us to prove lower bounds is to concentrate on proving that the symbolic distance measure lower bounds the PAA distance measure.

III. MATH

A. Dimensionality Reduction via PAA

To reduce the time series from n dimensions to, w dimensions, the data is divided into w equal sized frames. The mean value of the data falling within a frame is calculated and a vector of these values becomes the data-reduced representation.

The PAA dimensionality reduction is intuitive and simple, yet has been shown to rival more sophisticated dimensionality reduction techniques.

We normalize each time series to have mean of zero and a standard deviation of one before converting it to the PAA representation since it is well understood that it is meaningless to compare time series with different offsets and amplitudes.

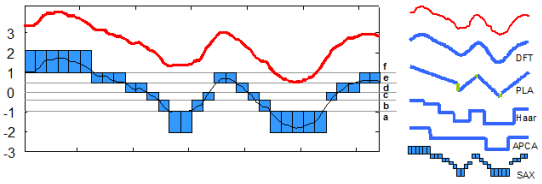
B. Discretization

- Having transformed a time series database into the PAA we can apply a further transformation to obtain a discrete representation. It is desirable to have a discretization technique that will produce symbols with equiprobability.

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j$$

PAA approximates a time series

- This is easily achieved since the normalized time series have a Gaussian distribution. To illustrate this, we extracted subsequences of length 128 from 8 different time series and plotted normal probability plots of the data as shown. A normal probability plot is a graphical technique that shows if the data is approximately normally distributed.



Visual Comparison of sequence of len 128 from 8 different time series

- As the figure shows, the highly linear nature of the plots suggests that the data is approximately normal. For a large family of the time series data in our disposal, we notice that the Gaussian assumption is indeed true.
- The correctness of the algorithm is guaranteed by the lower-bounding property of the distance measure in the symbolic space, which will be explained in the next section. Given that the normalized time series have highly Gaussian distribution, we can simply determine the breakpoints that will produce an equal-sized area under the Gaussian curve.
- Once the breakpoints have been obtained we can discretize a time series in the following manner. We first obtain a PAA of the time series. All PAA coefficients that are below the smallest breakpoint are mapped to the symbol a, all coefficients greater than or equal to the smallest breakpoint and less than the second the smallest breakpoint are mapped to the symbol b, etc.

C. Distance Measure

- Having introduced the new representation of time series, we can now define a distance measure on it. By far the most common distance measure for time series is the Euclidean distance. Given two-time series Q and C of the same length n, the image below defines their Euclidean distance, and image illustrates a visual intuition of the measure.

$$D_{PAA}(\bar{X}, \bar{Y}) \equiv \sqrt{\frac{n}{M}} \sqrt{\sum_{i=1}^{M} (\bar{x}_i - \bar{y}_i)}$$

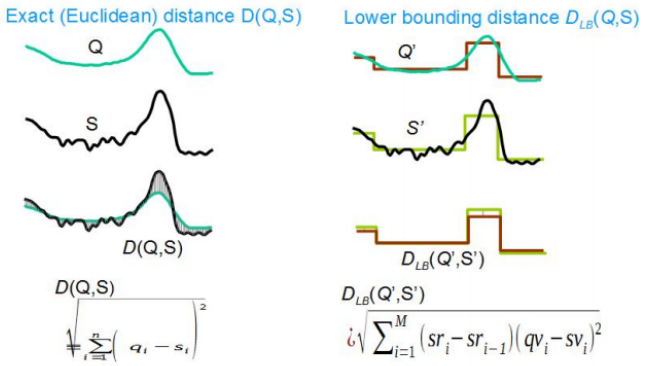
Reduced PAA equation

- If we transform the original subsequences into PAA representations, Q and C, using equation given, we can then obtain a lower bounding approximation of the Euclidean distance between the original subsequences.

$$\sum_{i=1}^{n} (q_i - c_i)^2 \geq n(\bar{Q} - \bar{C})^2 \geq n(dist(\hat{Q}, \hat{C}))^2$$

Distance Comparison

- This measure is illustrated in the image. If we further transform the data into the symbolic representation, we can define a MINDIST function that returns the minimum distance between the original time series of two words.
- We end this section with a visual comparison between SAX and the four most used representations in the literature survey. We can see that SAX preserves the general shape of the original time series.



Lower bounding means that for all Q and S, we have…

$D_{LB}(Q',S') \leq D(Q,S)$

Lower Bounding

- Note that since SAX is a symbolic representation, the alphabets can be stored as bits rather than doubles, which results in a considerable amount of space-saving.
- Therefore, SAX representation can afford to have higher dimensionality than the other real-valued approaches, while using less or the same amount of space.

IV. Applications of SAX

SAX has had a large impact in industry and academia. Below we summarize some of this work, without attempting to be exhaustive. We can broadly classify this work into those who have simply used the original formulation of SAX to solve a particular problem, and those who have attempted to extend or mitigate SAX in some way.

A. Query by Content (Indexing)

The majority of work on time series data mining appearing in the literature has addressed the problem of indexing time series for fast retrieval. Indeed, it is in this context that most of the representations enumerated. Dozens of papers have introduced techniques to do indexing with a symbolic

approach, but without exception, the answer set retrieved by these techniques can be very different to the answer set that would be retrieved by the true Euclidean distance. It is only by using a lower bounding technique that one can guarantee retrieving the full answer set, with no false dismissals.

B. Clustering

Clustering is one of the most common data mining tasks, being useful in its own right as an exploratory tool, and also as a sub-routine in more complex algorithms.Its not surprising that SAX can sometimes outperform the simple Euclidean distance, especially on noisy data, or data with shifting on the time-axis. More generally, we observed that SAX closely mimics Euclidean distance on various datasets.

C. Classification

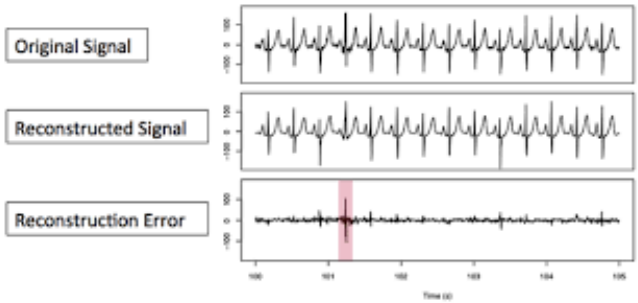
- Nearest Neighbour Classification: Once again, SAXs ability to beat Euclidean distance is probably due to the smoothing effect of dimensionality reduction, nevertheless, this experiment does show the superiority of SAX over the others proposed in the literature.
- Decision Tree Classification: Because of the Nearest Neighbors poor scalability, it is unsuitable for most data mining applications; instead, decision trees are the most common choice of the classifier. While decision trees are defined for real data, attempting to classify time series using the raw data would clearly be a mistake, since the high dimensionality and noise levels would result in a deep, bushy tree with poor accuracy.

D. Anomaly Detection

An Anomaly (outlier) is simply an unusual subsequence of the series. Unusual can be taken as improbable. Probability is not defined for the time series, Probability can be defined for Symbols. Mapping a time series to a symbol may allow us to assign a probability to the time series subsequence. This involves mapping the time series subsequence to a symbol in some symbol space. Key advantages of using SAX is that it yields an explainable model. Wherein the result of the model should not be underestimated.

Anomalies or outliers come in three types.

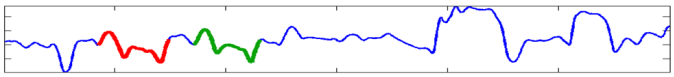
- Point Anomalies. If an individual data instance can be considered as anomalous with respect to the rest of the data (e.g. purchase with large transaction value)
- Contextual Anomalies, If a data instance is anomalous in a specific context, but not otherwise (anomaly if occur at a certain time or a certain region. e.g. large spike at the middle of the night)
- Collective Anomalies. If a collection of related data instances is anomalous with respect to the entire dataset, but not individual values. They have two variations.
  1. Events in unexpected order (ordered. e.g. breaking rhythm in ECG)
  2. Unexpected value combinations (unordered. e.g. buying a large number of expensive items)



Anomaly Detection in Time Series Data

E. Motif Discovery

In previous work, we defined the related concept of time series motif. Time series motifs are close analogs of their discrete cousins, although the definitions must be augmented to prevent certain degenerate solutions. The nave algorithm to discover the motifs is quadratic in the length of the time series. In, we demonstrated a simple technique to mitigate the quadratic complexity by a large constant factor, nevertheless this time complexity is clearly untenable for most real datasets.



Motif - Recurrent Pattern

F. Visualization

Data visualization techniques are very important for data analysis since the human eye has been frequently advocated as the ultimate data-mining tool. However, despite their illustrative nature, which can provide users better understanding of the data and intuitive interpretation of the mining results, there has been surprisingly little work on visualizing large time series datasets. One reason for this lack of interest is that time series data are also usually very massive in size. With limited pixel space and the typically enormous amount of data at hand, it is infeasible to display all the data on the screen at once, much less finding any useful information from the data. How to efficiently organize the data and present them in such a way that is intuitive and comprehensible to human eyes thus remains a great challenge

V. CONCLUSIONS

In this work, we have formulated the first dimensionality/numerosity reduction, lower bounding symbolic an approach in the literature. We have shown that our representation is competitive with, or superior to, other representations on a wide variety of classic data mining problems and that its discrete nature allows us to tackle emerging tasks such as anomaly detection, motif discovery, and visualization. A host of future directions suggest themselves. There is an enormous wealth of useful definitions, algorithms and data structures in the bioinformatics literature that can be exploited by our representation. It may be possible to create a lower bounding approximation of Dynamic Time Wrapping, by slightly

modifying the classic string edit distance. Finally, there may be utility in extending our work to multidimensional and streaming time series

ACKNOWLEDGMENT

The authors would like to thank Prof.SRS.Prabaharan for their constant support. We would like to thank Prof. B. Neppolian along with their entire team for organizing this event. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

REFERENCES

[1] Applying multiple time series data mining to large-scale network traffic analysis Weisong He,; Guangmin Hu,; Xingmiao Yao,; Guangyuan Kan,; Hong Wang,; Hongmei Xiang 2008

[2] Chiu, B. Keogh, E., Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. In the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24 - 27, 2003. Washington, DC, USA. pp 493-498.

[3] F. Duchene, C. Garbay, V. Rialle, "Mining heterogeneous multivariate time-series for learning meaningful patterns: Application to home health telecare," Research Report 1070-I, Institut d Informatique et Mathematiques Appliquees de Grenoble (IMAG), Grenoble, France, 2004.

[4] Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, Scott Sirowy (2006). Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating Systems. ICDM 2006.

[5] Li Wei, Eamonn Keogh and Xiaopeng Xi (2006) SAXually Explict Images: Finding Unusual Shapes. ICDM 2006.

[6] T. Armstrong and T .Oates. RIPTIDE: Segmenting Data Using Multiple Resolutions. In the Proceedings of the 6th IEEE International Conference on Development and Learning (ICDL), 2007.

[7] Motif Detection Inspired by Immune Memory (2007). William Wilson, Phil Birkin, and Uwe Aickelin.

[8] Heeyoul Choi, Chen Yu, Olaf Sporns and Linda Smith, "From Data Streams to Information Flow: Information Exchange in Child-Parent Interaction," The Annual Meeting of the Cognitive Science Society (CogSci 2011), Boston, MA. July 20-23, 2011.

[9] Voigtmann, C.; Lau, S. L. David, K. (2011), An Approach to Collaborative Context Prediction, in '2011 IEEE International Conference on Pervasive Computing and Communications Workshops. IEEE

[10] Improving the Classification Accuracy of Streaming Data Using SAX Similarity Features. Pekka Siirtola et al Pattern Recognition Letters.

[11] Legato and Glissando identification in Classical Guitar. Ozaslan and Arcos 2010. 7th Sound and Music Computing Conference Attack Based Articulation Analysis of Nylon String Guitar. Ozaslan and Arcos CMMR2010

[12] 3D Time-Varying Data Visualization Method Technique Featuring Symbolic Aggregate approximation ,M. Imoto, T. Itoh, IEEE Pacific Visualization 2011.

[13] RA-SAX: Resource-Aware Symbolic Aggregate approXimation for Mobile ECG Analysis, Hossein Tayebi , Shonali Krishnaswamy ,Agustinus Waluyo, Abhijat Sinha , , Mohamed Gaber.

[14] Relevant shape contour snippet extraction with metadata supported hidden Markov models. Wang and Candan. CIVR 10.

[15] A 3D Visualization Technique for Large Scale Time-Varying Data. Maiko Imoto, Takayuki Ito

[16] A New Symbolic Representation for the Identification of Informative Genes in Replicated Microarray Experiments. (2010) Jeremy D. Scheff, Richard R. Almon, Debra C. DuBois, William J. Jusko, and Ioannis P. Androulakis

[17] Relevant shape contour snippet extraction with metadata supported hidden Markov models. Wang and Candan. CIVR 10.

[18] Multiple Kernel Learning for Heterogeneous Anomaly Detection: Algorithm and Aviation Safety Case Study. Santanu Das, Bryan Matthews, Ashok Srivastava, ; Nikunj Oza, NASA Ames Research Center

[19] Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? Indre Zliobaite

[20] MAGIC 2.0: A Web Tool for False Positive Prediction and Prevention for Gesture Recognition Systems. Daniel Kohlsdorf, Thad Starner, Daniel Ashbrook: FG' 11, 2011