Symbolic Aggregate ApproXimation (SAX) under Interval Uncertainty

Chrysostomos D. Stylios[1] and Vladik Kreinovich[2]

[1]Laboratory of Knowledge and Intelligent Computing
Department of Computer Engineering
Technological Educational Institute of Epirus
47100 Kostakioi, Arta, Greece, stylios@teiep.gr

[2]Department of Computer Science
University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
vladik@utep.edu

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

1. Formulation of the Problem

- *Need for diagnostics:* often, we are monitoring a certain process for possible problems; e.g.:
  - we check the observed vibrations of a mechanical system indicate an abnormality;
  - we check the vital signs of a patient to see if an urgent medical intervention is needed.
- Sometimes, we have an algorithm that, based on the observations, decided whether intervention is needed.
- However, in most practical applications – especially in medicine – no such algorithm is readily available.
- What we have instead is numerous past data series corresponding both:
  - to cases when situation turned out to be normal,
  - and to cases with abnormality.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

2. Formulation of the Problem (cont-d)

- We have numerous past data series corresponding both:
  - to cases when situation turned out to be normal,
  - and to cases with abnormality.
- We thus need to extract such an algorithm from all these examples, i.e., use *machine learning*.
- Most machine learning algorithms work well if we have up to dozens of inputs.
- However, as a result of monitoring, we get values $x(t)$ corresponding to hundreds of moments of time t.
- So, to efficiently apply machine learning algorithms, we first need to compress the input data.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

3. Symbolic Aggregate approXimation (SAX): Main Idea

- The main objective of monitoring is to catch deviations from the normal regimes as early as possible.
- As a result, monitoring is performed at a high rate, to catch a deviation while this deviation is small.
- Thus, when the monitoring is arranged properly, values change very little from one moment to the next.
- So, we can safely replace the original function $x(t)$ with a piece-wise constant approximation.
- On each interval, we store only its endpoints and the value of the function on this interval.
- This representation indeed leads to a drastic reduction in data size.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

4. Symbolic Aggregate approXimation (cont-d)

- A further compression is possible since:
  - a computer-represented real number require dozens of bits to store, corresponding to ten decimal digits,
  - but measurements accuracy is usually 1–10%, so two decimal digits are enough.
- Symbolic Aggregate approXimation (SAX) is a technique for such a reduction.
- In the interval $[\underline{x}, \overline{x}]$ of possible values of $x(t)$, we select thresholds $x_0 = \underline{x}, x_1, x_2, \ldots, x_m$.
- Then, for each moment of time t, instead of storing $x(t)$, we store the index i for which $x(t) \in [x_i, x_{i+1}]$.
- At present, SAX is the most efficient data compression technique.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

5. SAX: Details and Successes

- To maximize the amount of information after compression, SAX takes into account that:
  - the maximum amount of Shannon's information $-\sum\limits_{i=0}^{m} p_i\cdot\log_2(p_i)$, where $p_i=\text{Prob}(x(t)\in[x_i,x_{i+1}])$,
  - is attained when all the probabilities p_i are equal to each other – and is, thus, equal to $p_i=\dfrac{1}{m+1}$.
- Thus, SAX selects the thresholds x_i for which

$$p_i=\text{Prob}(x(t)\in[x_i,x_{i+1}])=\frac{1}{m+1}.$$

- SAX techniques led to many practical applications ranging from engineering to medicine.

6. SAX: Problem

- Measurement errors were a motivation for SAX techniques.
- However, SAX does not take measurement errors into account.
- So, we often get thresholds x_i and x_{i+1} which are much closer to each other than the measurement accuracy.
- Sometimes, x_i and x_{i+1} differ by 5% while the measurement accuracy is 10%.
- In this case, we cannot tell whether the actual value $x(t)$ was in the i-th interval or in the next interval.
- It is therefore desirable to explicitly take measurement uncertainty into account in SAX techniques.
- This is what we do in this paper.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

7. Case When Measurement Inaccuracy Can Be Ignored (Reminder)

- Based on the observed values $x(t)$, we can find the probabilities with which different values of x occur.
- These probabilities can be naturally described by a probability density function $\rho(x)$, with $\int \rho(x)\,dx = 1$.
- In many practical situations, the observed signal is a joint effect of many different independent processes.
- In such situations, the Central Limit Theorem implies that the resulting distribution is Gaussian.
- We want to select the thresholds $x_1, x_2, \ldots$
- We can describe, for every value x, the number $\rho_t(x)$ of thresholds per unit length; the total is $\int \rho_t(x)\,dx = m$.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

8. Case of No Measurement Inaccuracy (cont-d)

- After the data compression, the only information that we have about each value $x(t)$ in the index i.
- So, to reconstruct the value $x(t)$ based on this information, we select the midpoint $\widetilde{x}(t)$ of the i-th subinterval.
- This reconstruction is approximate, there is an approximation error $\varepsilon(t) \stackrel{\text{def}}{=} \widetilde{x}(t) - x(t) \neq 0$.
- Ideally, we would like to have all these errors to be as close to 0 as possible.
- The vector $\varepsilon = (\varepsilon(t_1), \varepsilon(t_2), \ldots)$ of these errors should be close to the zero vector $\vec{0} = (0, 0, \ldots)$:

$$d(\varepsilon, \vec{0}) = \sqrt{\sum_k (\varepsilon(t_k))^2} \to \min .$$

- In the continuous approximation, this is equivalent to minimizing $\int (\varepsilon(t))^2 \, dt$.

Formulation of the ...
Symbolic Aggregate ...
SAX: Problem
Towards Formulating ...
Case of Interval ...
How Measurement ...
How Measurement ...
Solving the ...
What If We Minimize ...

9. Alternative Ideas

- The least-squares approach is vulnerable to outliers.
- The second idea is to avoid this sensitivity by using ℓ^p-estimates:
$$\int |\varepsilon(t)|^p \, dt \to \min .$$
- The third idea is to explicitly minimize the number of bits needed to describe all the thresholds.
- If $x_{i+1} - x_i \approx 2^{-b}$, then it is sufficient to describe the first b binary digits of the corresponding interval.
- This, the number of bits needed to store each threshold is approximately equal to $b \approx -\log_2(x_{i+1} - x_i)$.
- So, we minimize the average number of bits, i.e., the sum $-\sum\limits_k \log_2(x_{i+1} - x_i)$ or the corresponding integral.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

10. Towards Formulating the Corresponding Optimization Problems in Precise Terms

- On the unit interval I around a value x, there are $\rho_t(x)$ thresholds.
- Thus, I is divided into $\rho_t(x)$ subintervals.
- Hence, the width $w = x_{i+1} - x_i$ of each subinterval can be estimated as the ratio $w = \dfrac{1}{\rho_t(x)}$.
- The absolute value $a \stackrel{\text{def}}{=} |\varepsilon|$ of $\varepsilon \stackrel{\text{def}}{=} x_{\text{mid}} - x$ is uniformly distributed on $\left[0, \dfrac{w}{2}\right] = \left[0, \dfrac{1}{2\rho_t(x)}\right]$.
- This uniform distribution has a probability density

$$\rho_0(a) = \frac{1}{w/2} = \frac{2}{w}.$$

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

11. Formulating the Problem (cont-d)

- The average value of ε^2 on this interval equals

$$\int_0^{w/2} a^2 \cdot \rho_0(a)\, da = \frac{2}{9} \cdot w^2 = \text{const} \cdot \frac{1}{(\rho_t(x))^2}.$$

- Each value x occurs with probability density $\rho(x)$.
- So, minimizing the integral $\int (\varepsilon(t))^2\, dt$ is equivalent to minimizing the integral $\int \rho(x) \cdot \dfrac{1}{(\rho_t(x))^2}\, dx$.
- Similarly, minimizing $\int |\varepsilon(t)|^p\, dt$ is equivalent to minimizing the integral $\int \rho(x) \cdot \dfrac{1}{(\rho_t(x))^p}\, dx$.
- For minimizing the number of bits, for each interval, $x_{i+1} - x_i \approx \dfrac{1}{\rho_t(x)}$.
- So, $-\log_2(x_{i+1} - x_i) = -\text{const} \cdot \ln(\rho_t(x))$, and we need to minimize the integral $-\int \rho(x) \cdot \ln(\rho_t(x))\, dx$.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

12. Solving the Optimization Problems

- For the least squares optimization, we minimize $\int \rho(x) \cdot \dfrac{1}{(\rho_t(x))^2}\,dx$ under constraint $\int \rho_t(x)\,dx = m$.
- Lagrange multiplier method mean optimizing

$$\int \rho(x) \cdot \frac{1}{(\rho_t(x))^2}\,dx + \lambda \cdot \int \rho_t(x)\,dx;$$

  - differentiating this objective function with respect to each unknown $\rho_t(x)$ and
  - equating the resulting derivative to 0,
  - we conclude that $-2 \cdot \dfrac{\rho(x)}{(\rho_t(x))^3} + \lambda = 0$,
  - so $\rho_t(x) = \text{const} \cdot (\rho(x))^{1/3}$.
- The corresponding constant can be found from the condition $\int \rho_t(x)\,dx = m$, so $\rho_t(x) = \dfrac{(\rho(x))^{1/3}}{\int (\rho(y))^{1/3}\,dy}$.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

13. Solving the Optimization Problems (cont-d)

- *Example:* $\rho(x)$ is normally distributed, with mean μ and variance σ^2.
- *Solution:* $\rho_t(x)$ is also normal, with the same mean and the variance $\dfrac{\sigma^2}{3}$.
- ℓ^p*-case:* we get $\rho_t(x) = \dfrac{(\rho(x))^{1/(p+1)}}{\int (\rho(y))^{1/(p+1)}\,dy}$.
- For normal $\rho(x)$, the distribution $\rho_t(x)$ is also normal, with the same mean and the variance $\dfrac{\sigma^2}{p+1}$.
- *For the bit minimization:* we get $\rho_t(x) = m \cdot \rho(x)$.
- Here, the probability p_i of being in a subinterval is the same for all the subintervals i.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

14. Case of Interval Uncertainty

- In the ideal world:
  - for each measuring instrument,
  - we should know the probability distribution of measurement errors.
- This distribution can be determined if we compare:
  - the results of the given measuring instrument
  - with the results of a super-precise "standard" measuring instrument.
- This "calibration" process is possible, but it is usually very costly.
- Indeed, ensors are cheap nowadays, but super-precise measuring instruments are not.
- As a result, in many cases, all we know is the upper bound Δ on the absolute measurement error.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

15. How Measurement Errors Affect Threshold Selection

- In the ideal case, any deviation of the midpoint $\widetilde{x}(t)$ from the actual signal $x(t)$ is an inaccuracy.
- However, if we take measurement uncertainty into account, then deviations not exceeding Δ are OK.
- Indeed, the (unknown) actual value of the measured quantity can be anywhere within $[x(t)-\Delta, x(t)+\Delta]$.
- So, if $\widetilde{x}(t)$ is within this interval, it can still be exactly equal to the actual value.
- Only when $|\varepsilon(t)| > \Delta$, we know that there is an approximation error.
- This error can be gauged as the distance

$$d(\widetilde{x}(t), [x(t)-\Delta, x(t)+\Delta]) =$$
$$\min\{d(\widetilde{x}(t), x) : x \in [x(t)-\Delta, x(t)+\Delta]\}.$$

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

16. How Measurement Errors Affect Threshold Selection (cont-d)

- One can check that this distance is equal to

$$d(\widetilde{x}(t), [x(t)-\Delta, x(t)+\Delta]) = \max(|\varepsilon(t)|-\Delta, 0).$$

- This distance is what we should take into account (instead of $|\varepsilon(t)|$) when we optimize.
- The average value of the square of the distance is:

$$\frac{2}{w}\cdot\int_0^{w/2}(\max(a-\Delta,0))^2\,da.$$

- So, in the least square cases, we minimize:

$$\int\left(\frac{1}{\rho_t(x)}-2\Delta\right)^3\cdot\rho_t(x)\cdot\rho(x)\,dx.$$

- For the p-th powers, we similarly minimize:

$$\int\left(\frac{1}{\rho_t(x)}-2\Delta\right)^{p+1}\cdot\rho_t(x)\cdot\rho(x)\,dx.$$

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

17. Solving the Optimization Problems

- To solve these constraint optimization problems, we:
  - apply the Lagrange multiplier methods to reduce them to unconstraint optimization, and
  - equate derivatives to 0.
- For the Least Squares cases, we get the equation

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^2 \cdot \left(\frac{2}{\rho_t(x)} + 2\Delta\right) = \frac{\lambda}{\rho(x)}.$$

- This is a cubic equation in terms of the unknown $\dfrac{1}{\rho_t(x)}$.
- For the ℓ^p-case, we get the equation

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^p \cdot \left(\frac{p}{\rho_t(x)} + 2\Delta\right) = \frac{\lambda}{\rho(x)}.$$

- The parameter λ can be determined from the condition $\int \rho_t(x)\,dx = m$.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

18. What If We Minimize the Number of Bits

- In this case, the only restriction is that the width $w = \frac{1}{\rho_t(x)}$ cannot be smaller than 2Δ.
- Thus, the threshold density $\rho_t(x)$ cannot be larger than $\frac{1}{2\Delta}$.
- Minimizing the number of bits under this constraint leads to

$$\rho_t(x) = C \cdot \min\left(\rho(x), \frac{1}{2\Delta}\right).$$

- The constant C must also be determined from the condition that $\int \rho_t(x)\,dx = m$.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

19. Conclusions

- Symbolic Aggregate Approximations (SAX) is a technique for data compression.
- The intent of SAX is to take uncertainty into account.
- However, the current implementations of SAX do not account for all the uncertainty.
- So, we propose to extend the current SAX methodology to taking interval uncertainty into account.
- Specifically, we propose to take interval uncertainty into account when selecting the thresholds.
- In this talk, we propose theoretical foundations and the resulting asymptotically optimal algorithms.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

20. Future Work

- It is desirable to test the new algorithms on several real-life examples.
- The new algorithms lead to an asymptotically better data compression.
- This will hopefully lead to faster computations.
- However, implementing these algorithms requires an additional computational overhead.
- We know that asymptotically, the advantages outweigh this overhead.
- Testing on real-life examples would help us:
  - to check whether the new algorithm is still beneficial for real-size data,
  - and if this is not always the case, to find out when the new algorithm should be recommended.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .

21. Acknowledgment

- This work was supported in part by the National Science Foundation grants:
  - HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
  - DUE-0926721.
- This work was performed when C. Stylios was a Visiting Researcher at the University of Texas at El Paso.

Formulation of the . . .
Symbolic Aggregate . . .
SAX: Problem
Towards Formulating . . .
Case of Interval . . .
How Measurement . . .
How Measurement . . .
Solving the . . .
What If We Minimize . . .