Intuitionistic Fuzzy Estimations for Similarity Queries Using Sketches of Numeric Data

Boyan Kolev

INSTITUTE OF MATHEMATICS AND INFORMATICS
BULGARIAN ACADEMY OF SCIENCES

Vassia Atanassova
Peter Vassilev

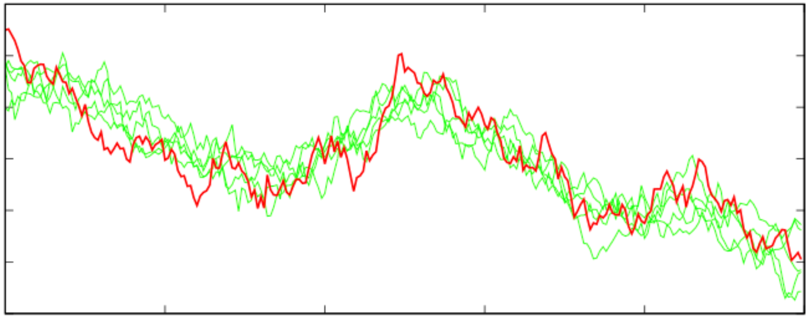
IBPhBME

Twenty Second International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets
Warsaw, Poland, October 18, 2024

Work Overview

Context

- Consider similarity measures over time series
  - Represented as multi-dimensional numerical vectors



- Sketching for dimensionality reduction
  - Project vectors onto a space of much lower dimensionality
  - Approximately preserve Euclidean distances
  - Lower response time ←→ small accuracy loss

Work Overview

Objectives

- Intuitionistic fuzzy estimation of similarity measures
  - "is *x* similar to *q*?"
  - Truth mapped to similarity itself
  - Indefiniteness mapped to loss of accuracy
- Validation through SQL-like queries
  - Result set is intuitionistic fuzzy
  - Each row is assigned degrees of membership and non-membership

```
SELECT x
FROM tsdata
WHERE x is similar to $q
```

Random Projection

Sketching for Dimensionality Reduction

- Sketch computation

time series:
n-dimensional row vectors

transformation matrix:
n x d
$r_{i,j} = \{-1 \mid 1\}$

sketches:
d-dimensional
d << n

$\frac{1}{\sqrt{d}}$

t_1 =	$t_{1,1}$ … … … $t_{1,n}$
t_2 =	$t_{2,1}$ … … … $t_{2,n}$
……	… … …
t_m =	$t_{m,1}$ … … … $t_{m,n}$

X

r_1 =	**r_2 =**	**…**	**r_d =**
$r_{1,1}$	$r_{2,1}$	…	$r_{d,1}$
$r_{1,2}$	$r_{2,2}$	…	$r_{d,2}$
…	…	…	…
$r_{1,n}$	$r_{2,n}$	…	$r_{d,n}$

=

s_1 =	$s_{1,1}$ … $s_{1,d}$
s_2 =	$s_{2,1}$ … $s_{2,d}$
……	…
s_m =	$s_{m,1}$ … $s_{m,d}$

Random Projection

Control Parameters

*D. Achlioptas, **Database-friendly Random Projectio***
ACM SIGMOD-PODS, May, Santa Barbara, CA, 2001

Given:

ε – accuracy control parameter
β – probability control parameter
m – number of time series
n – time series dimension
d – reduced dimension

If $$d \geq \frac{4 + 2\beta}{\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}} \log(m)$$

Then, with probability $Pr = 1 - m^{-\beta}$

Sketches preserve the distance with distortion ε:

$$(1-\varepsilon)\left\|t_i - t_j\right\|^2 \leq \left\|s_i - s_j\right\|^2 \leq (1+\varepsilon)\left\|t_i - t_j\right\|^2$$

- Q: Is *x* similar to *y*?
  - Squared distance

$$D_{x,y} = \|x - y\|^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$

  - Pearson correlation

$$C_{x,y} = \frac{1}{S_x S_y} \sum_{i=1}^{n} (x_i - M_x)(y_i - M_y)$$

    - to convert from distance

$$C_{x,y} = \frac{1}{S_x S_y} \left(\frac{Q_x + Q_y - D_{x,y}}{2} - n M_x M_y \right)$$

  - Degree of truth: C scaled to [0, 1]

$$T_{x,y} = \frac{1 + C_{x,y}}{2}$$

TS summary:

$$M_x = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$S_x = \sqrt{\sum_{i=1}^{n} (x_i - M_x)}$$

$$Q_x = \sum_{i=1}^{n} {x_i}^2$$

Intuitionistic Fuzzy Estimates

of Similarity Measures

- Q: Is *x* similar to *y*?
  - For efficiency: compute similarity on sketches *u* and *v*
  - Accuracy loss → degree of indefiniteness

$$D_{x,y}^{min} = \frac{D_{u,v}}{1+\varepsilon} \leq D_{x,y} \leq \frac{D_{u,v}}{1-\varepsilon} = D_{x,y}^{max}$$

$$C_{x,y}^{min} = \frac{1}{S_x S_y}\left(\frac{Q_x + Q_y - D_{x,y}^{max}}{2} - nM_x M_y\right) \qquad C_{x,y}^{max} = \frac{1}{S_x S_y}\left(\frac{Q_x + Q_y - D_{x,y}^{min}}{2} - nM_x M_y\right)$$

- A: $(\mu_{x,y}, \nu_{x,y})$:

$$\mu_{x,y} = \frac{1 + C_{x,y}^{min}}{2} \qquad \nu_{x,y} = \frac{1 - C_{x,y}^{max}}{2}$$

Validation

Experimental setup

Parameters:

ε (accuracy control parameter) = 0.2
β (probability control parameter) = 1
m (number of time series) = 100
n (time series dimension) = 100000
d (reduced dimension) = 1600
Pr = 0.99

PostgreSQL schema

```
CREATE TABLE data (
    id integer primary key,
    x double precision[],    -- original vector, len=100000
    mean double precision,   -- Mx, i.e. mean of x
    nstd double precision,   -- Sx, i.e. stddev w/o division by n
    sum2 double precision,   -- Qx, i.e. sum of squares
    sk double precision[]    -- sketch, len=1600
);
```

Validation

- Data preparation process
  - Generate time series synthetically from N(0, 1)
  - Generate a random matrix into a table:

```
CREATE TABLE rndmx (
    vec boolean[]    -- random vector, len=100000, {t/f} → {±1}
);                   -- contains 1600 rows
```

  - Compute sketches and store into column: ***sk***
  - Compute summaries into the additional columns:
    - *mean* (M_x), *nstd* (S_x), *sum2*(Q_x)

Validation

- Query on original time series
  - returns degree of membership

```
PREPARE sim_ts AS
select id, (1+c)/2 as mu from (
    select id, ((qsum2+dsum2-d*d)/2 - n*qmean*dmean)/(qnstd*dnstd) as c
    from (select data.id as id,
            array_length(qs.x, 1) as n, array_dist(qs.x, data.x) as d,
            qs.nstd as qnstd, qs.mean as qmean, qs.sum2 as qsum2,
            data.nstd as dnstd, data.mean as dmean, data.sum2 as dsum2
        from data qs, data
        where qs.id = $1
    ) as t
) as t;
```

Validation

- Intuitionistic fuzzy query on sketches
  - returns degrees of membership and non-membership

```
PREPARE sim_sketch AS
select id, (1+cmin)/2 as mu, (1-cmax)/2 as nu from (
    select id, ((qsum2+dsum2-d*d/0.8)/2 - n*qmean*dmean)/(qnstd*dnstd) as cmin
             , ((qsum2+dsum2-d*d/1.2)/2 - n*qmean*dmean)/(qnstd*dnstd) as cmax
    from (select data.id as id,
            array_length(qs.x, 1) as n, array_dist(qs.sk, data.sk) as d,
            qs.nstd as qnstd, qs.mean as qmean, qs.sum2 as qsum2,
            data.nstd as dnstd, data.mean as dmean, data.sum2 as dsum2
        from data qs, data
        where qs.id = $1
    ) as t
) as t;
```

Validation

- Performance improvement
  - 46 times in our setup (dim. n=100000 → d=1600)

```
ifsk2=# execute sim_ts(1);
Time: 6523.801 ms (00:06.524)
ifsk2=# execute sim_sketch(1);
Time: 141.545 ms
```

```
id |          mu          |          nu
----+----------------------+-----------------------
  1 |   0.9999999999999998 | 2.220446049250313e-16
  2 |   0.3693058085968354 |    0.420461627789663
  3 |   0.3340468220549211 |    0.443965380508609
  4 | 0.38293512934554136 |    0.4113762596527037
  5 | 0.36981105610112064 |    0.4201249993652301
  6 |   0.3383522120735153 |   0.44109817313041266
```

- Conclusion
  - Accuracy loss mapped to intuitionistic fuzzy uncertainty
  - Can be combined with other intuitionistic fuzzy estimates
    - e.g. predicates in the same IFSQL query *

* Kolev B. ***Intuitionistic Fuzzy PostgreSQL***. Advanced Studies in Contemporary Mathematics, Vol. 11, No. 2, 2005. 163-177

Kolev, B. ***Intuitionistic Fuzzy Relational Databases And Translation of the Intuitionistic Fuzzy SQL***. Proceeding of the Sixth International FLINS Conference, 1-3 Sep 2004, Blankenberge, Belgium. 189-194

Thank you for the attention!

Boyan Kolev

bkolev@math.bas.bg

INSTITUTE OF MATHEMATICS AND INFORMATICS
BULGARIAN ACADEMY OF SCIENCES

The second and the third authors are grateful for the support under Bulgarian National Science Fund, Grant No. KP-06-N-72/8 from 14.12.2023.