Learning experts' preferences from informetric data

Marek Gagolewski[1,2] **Jan Lasek**[3]

[1]Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland, gagolews@ibspan.waw.pl

[2]Faculty of Mathematics and Information Science, Warsaw University of Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland

[3]Interdisciplinary PhD Studies Program,
Institute of Computer Science, Polish Academy of Sciences
j.lasek@phd.ipipan.waw.pl

IFSA-EUSFLAT
Gijon, Spain, 2015

Introduction and motivation

stackoverflow

787,720 REPUTATION

● 363 ● 5417 ● 6629

Jon Skeet top 0.01% overall

- 6313 Why is subtracting these two times (in 1927) giving a strange r...
- 1852 What's the difference between String and string?
- 1801 Why is char[] preferred over String for passwords?
- 1285 Difference between Decimal, Float and Double in .NET?
- 1172 What are the correct version numbers for C#?

597,277 REPUTATION

● 100 ● 2042 ● 2072

Darin Dimitrov top 0.01% overall

- 807 File Upload ASP.NET MVC 3.0
- 393 How to send a PUT/DELETE request in jQuery?
- 355 How do I specify different Layouts in the ASP.NET MVC 3 ra
- 326 Using Ajax.BeginForm with ASP.NET MVC 3 Razor
- 259 ASP.NET MVC3 - textarea with @Html.EditorFor

Introduction and motivation

Google Scholar

Lotfi A. Zadeh

Follow

Professor Emeritus, EECS, UC Berkeley
Fuzzy Logic, Soft Computing, Artificial Intelligence, Human-Level Machine Intelligence
Verified email at eecs.berkeley.edu - Homepage

Title 1–20	Cited by	Year
Fuzzy Sets L Zadeh Information and Control 8, 338-353	57186	1965
Outline of a new approach to the analysis of complex systems and decision processes LA Zadeh Systems, Man and Cybernetics, IEEE Transactions on, 28-44	17370	1973
The concept of a linguistic variable and its application to approximate reasoning—I LA Zadeh Information sciences 8 (3), 199-249	10900	1975
Decision making in a fuzzy environment		

Introduction and motivation

The field of *informetrics* deals with measurable aspects of information processes. So far, a number of tools has been suggested to quantify the value of information.

In this exposition we will investigate the efficacy of a set of chosen off-the-shelf solutions in an exemplary setup.

Producer Assessment Problem (PAP)

Let us formally definite the problem under our consideration [Gagolewski and Grzegorzewski 2011].

Producer Assessment Problem

Let $P = \{p_1, \ldots, p_k\}$ be a finite set consisting of k producers. The i-th producer outputs n_i products. Additionally, each product is given some kind of quantitative rating, e.g. concerning its overall quality.
The state of p_i may be described by a sequence

$$\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_{n_i}^{(i)}\right) \in \mathbb{I}^{1,2,\ldots} = \bigcup_{n \geqslant 1} \mathbb{I}^n$$

with elements in $\mathbb{I}$, e.g. $\mathbb{I} = [0, \infty)$. Most importantly, we should note that the numbers of products may vary from producer to producer. The **goal** is to **design tools for producers' evaluation (rankings, preference relations, etc.)** and **their impact measurement**.

Producer Assessment Problem visually

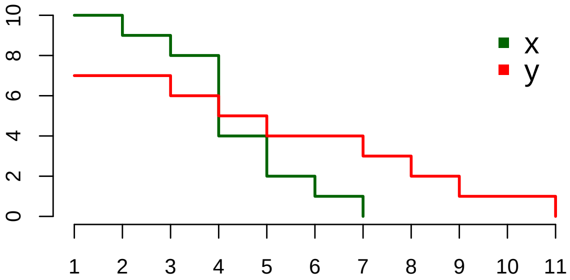


Figure: Illustration of PAP definition for two example output vectors $\mathbf{x} = (10, 9, 8, 4, 2, 1)$ and $\mathbf{y} = (7, 7, 6, 5, 4, 4, 3, 2, 1, 1)$.

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's *h*-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \ldots, i\}$ [Woeginger 2008].

.

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \ldots, i\}$ [Woeginger 2008].
- mean quality of a product $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \ldots, i\}$ [Woeginger 2008].
- mean quality of a product $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
- sum of all qualities $\Sigma(\mathbf{x}) = \sum_{i=1}^{n} x_i$,

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \ldots, i\}$ [Woeginger 2008].
- mean quality of a product $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
- sum of all qualities $\Sigma(\mathbf{x}) = \sum_{i=1}^{n} x_i$,
- maximal quality of a product x_1,

.

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \dots, i\}$ [Woeginger 2008].
- mean quality of a product $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i$,
- sum of all qualities $\Sigma(\mathbf{x}) = \sum_{i=1}^{n} x_i$,
- maximal quality of a product x_1,
- number of products n.

Available tools for analysis (1)

Up until today, a number of tools were proposed for Producer Assessment Problem including:

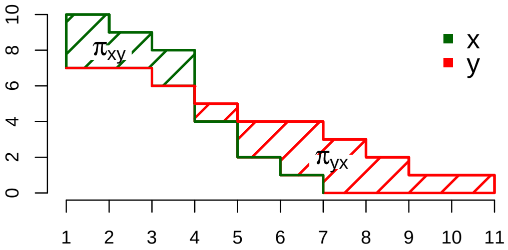
- Hirsch's h-index $i_H = \max\{i : x_i \geqslant i\}$ [Hirsch 2005],
- Egghe's g-index $i_G = \max\{i : \sum_{j=1}^{i} x_i \geqslant i^2\}$ [Egghe 2010],
- Woeginger's w-index $i_W = \max\{i : x_j \geqslant i - j + 1 \text{ for all } j = 1, \ldots, i\}$ [Woeginger 2008].
- mean quality of a product $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i$,
- sum of all qualities $\Sigma(\mathbf{x}) = \sum_{i=1}^{n} x_i$,
- maximal quality of a product x_1,
- number of products n.

These are examples of so–called **impact indexes**.

Available tools for analysis (2)

There are also tools from the domain of fuzzy systems. For example, the following fuzzy preference relation was suggested [Gagolewski and Lasek 2015]. For two output vectors **x** and **y**, the membership function of fuzzy preference relation $\mathbf{x} \blacktriangleleft \mathbf{y}$ is given by

$$\mu(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{\pi_{yx}}{\pi_{xy}+\pi_{yx}} & \text{if } \pi_{xy} + \pi_{yx} > 0, \\ 0.5 & \text{otherwise,} \end{cases}$$



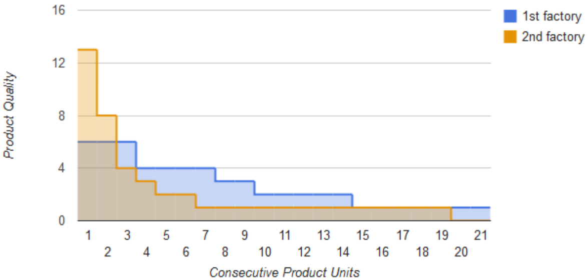
Research question

In this exposition, our research question is **which of the proposed functions (if any) is effective in describing experts' preferences in** an exemplary instance of Producer Assessment Problem. In other words: **¿Do these tools effectively compress information contained in data?**

We prepared generated data for PAP for an on–line questionnaire. The participants' (experts') responses serve us as evidence for validation purposes.

Questionnaire

In the questionnaire, participants were asked to provide answers for a series of questions.



Quality of output of the 1st factory (in total 21 units):

6 6 6 4 4 4 4 3 3 2 2 2 2 2 1 1 1 1 1 1 1

Quality of output of the 2nd factory (in total 19 units):

13 8 4 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1

Validation of hypothesis

To validate the hypothesis we confront two approaches:

- compare vectors on each coordinate and equalize their lengths by padding the shorter ones with zeros

$$(x_1, x_2, \cdots, x_n) \to (x_1, x_2, \ldots, x_n, 0, 0, \ldots, 0)$$

- extract certain features of output vectors using the discussed tools

$$(x_1, x_2, \cdots, x_n) \to (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x}))$$

We use several prediction models:

- Ordinal Logistic Regression,
- *k*-Nearest Neighbours classifier and
- Random Forest model.

The models are trained on 80% of data and evaluated on 20% ($\approx$ 1000 instances). In consecutive slides we discuss evaluation metrics used.

Evaluation metrics (1)

For ith example in the data set, $i = 1, 2, \dots, N$ let

- $l_t^{(i)}$, $t \in \{-2, -1, 0, 1, 2\}$ denote true preference label,
- a given model assign probability $\mathbb{P}(l_k^{(i)})$ to label l_k,
- a classifier assign labels according to $\hat{l}_p^{(i)} = \mathrm{argmax}_k\, \mathbb{P}(l_k^{(i)})$.

We considered the following evaluation measures:

- Missclassification rate

$$Misscl = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(l_t^{(i)} \neq \hat{l}_p^{(i)}).$$

- average distance between labels for $d(l_t^{(i)}, l_p^{(i)}) = |t - p|$

$$AvgDist = \frac{1}{N} \sum_{i=1}^{N} d(l_t^{(i)}, \hat{l}_p^{(i)}).$$

Evaluation metrics (2)

- Rank Probability Score

$$RPS = \frac{1}{4N} \sum_{i=1}^{N} \sum_{j=-2}^{2} \left(\hat{F}^{(i)}(j) - F^{(i)}(j) \right)^2,$$

  with $\hat{F}^{(i)}(\cdot)$ and $F^{(i)}(\cdot)$ being observed and estimated cumulative distribution function for labels

- Concordance Index

$$C = \frac{1}{M} \sum_{i:\ l_t^{(i)} \neq 0} \mathbb{1}(l_t^{(i)},\ \hat{l}_p^{(i)}\ concordant) + 0.5 \cdot \mathbb{1}(\hat{l}_p^{(i)} = 0).$$

  with M being the number of "usable pairs" (i.e., $l_t^{(i)} \neq 0$)

Results - evaluation of models

Below we present the results of experiment for the two approaches (marked with superscript i and c for the "index" and "coordinate" approach respectively).

Table: Results of classification.

	Misscl	*AvgDist*	*RPS*	C'
OLR_i	0.409	0.465	0.086	0.08
OLR_c	0.394	0.454	0.082*	0.075
$k\mathrm{NN}_i$	0.401*	0.457*	0.085*	0.083*
$k\mathrm{NN}_c$	0.453	0.548	0.099	0.122
RF_i	0.385*	0.452*	0.076*	0.078
RF_c	0.434	0.537	0.094	0.092
Equal	0.865	1.255	0.202	0.5

Results - feature importance (1)

For OLR model and different versions of kNN model (for various values of parameter k) we calculated how many times a given feature was picked by the employed feature selection procedure for different respondents. In this way, we obtain that the most important for classification are:

1. i_G (picked 42 times)
2. $\Sigma(\mathbf{x})$ (30)
3. $\bar{\mathbf{x}}$ (27)
4. x_1 (16)
5. FP (15)

Results - feature importance (2)

In case of Random Forest model we derived ranking of features aggregating individual importance rankings for 32 participants by Borda count. The following ranking of features was obtained (top 5):

1. FP
2. i_G
3. $\Sigma(\mathbf{x})$
4. $\bar{\mathbf{x}}$
5. x_1

Summary of results

Our findings can be summarized in the following points:

Summary of results

Our findings can be summarized in the following points:

- The emphasis was put on quality rather than productivity during the evaluation process.

Summary of results

Our findings can be summarized in the following points:

- The emphasis was put on quality rather than productivity during the evaluation process.
- The available tools are effective in compressing information from producers' output vectors.

Summary of results

Our findings can be summarized in the following points:

- The emphasis was put on quality rather than productivity during the evaluation process.
- The available tools are effective in compressing information from producers' output vectors.
- Among the best performing aggregation tools in our experiment we identified Egghe's g-index i_G, sum of product qualities $\Sigma(\mathbf{x})$, the fuzzy preference relation FP, mean quality of a product $\bar{\mathbf{x}}$ and the maximal quality of a product x_1.

Thank you for your attention!