Abstract
Distribution Extents (DE) of order \( k \) for a sample \( \{x_1, x_2, \ldots, x_n\} \) of a non-negative stochastic variable \( X \) can be defined as

\[
E_k = \begin{cases} 
  \left( \sum_{i=1}^{n} p_i^{k} \right)^{1/k} & \text{if } k \neq 1, \\
  \exp(-\sum_{i=1}^{n} p_i \ln p_i) & \text{if } k = 1
\end{cases}
\]

where \( p_i = x_i / \sum_{i=1}^{n} x_i \),

and are useful measures for a number of “large values” in the sample. They were introduced by Timo Koski and Lars-Erik Persson in 2003 [1], who generalized results of L.L. Campbell [2], and are generalization of inverse Herfindahl-Hirschman Index (HHI), a commonly accepted measure of market concentration in economics, Simpson’s diversity index used in ecology and are closely related to Shannon-Wiener Index and the Rényi entropy and divergence.

In this work we describe general properties of \( E_k \) and use it in analysis of a web advertisement network, where actors are advertisers, publishers, and users, for three purposes: 1) as cut off parameters to present the network as a graph to visualize the network and to use graph theory methods 2) as independent variable in predictive modeling, and 3) as a criterion for optimization of some parameters of models.

Keywords: distribution; entropy; web advertisement network; predictive modeling.

1. Introduction

A typical question in many data mining tasks is:
We have \( n \) objects, characterized by a variable \( x_i, i = 1, 2, \ldots n \) and number \( n \) is too large to consider all objects.

For simplicity we suppose that \( x_1 \geq x_2 \geq \cdots x_n \). We need truncate data dropping small objects and keeping in analysis only large object. What could be criteria to choose a cut-off parameter \( x_0 \) or threshold for \( x_i \), if we are going to keep only \( x_i > x_0 \)? How many large objects do we have?

Examples:
Countries: How many countries with large population? With large GDP?
Enterprises: In wide use is inverse Herfindahl-Hirschman Index (HHI), a commonly accepted measure of market concentration in economics[3]. In this case \( x_i \) - revenue, \( p_i = \frac{x_i}{\sum x_i} \) - market shares:

\[ HHII = \text{sum}(p_i^2); \]
InverseHHI = \frac{1}{\text{sum}(p_i^2)} \quad (1)

2. Properties of Inverse Herfindahl-Hirschman Index (HHI)

The inverse HHI has following practically important properties:

\[ 1 \leq \text{InverseHHI} \leq n; \]

\[ \text{InverseHHI} = 1, \quad \text{when} \quad x_1 = A, \ x_i = 0, \quad i = 2, \ldots n, \ A > 0 \quad (1a) \]

\[ \text{InverseHHI} = n, \quad \text{when} \quad x_1 = x_2 = \cdots = x_n \quad (1b) \]

Let \[ x_i = A, \quad i = 1, 2, \ldots n_{\text{large}}, \]

\[ x_i = a, \quad i = n_{\text{large}} + 1, \ldots n; \ \text{where} \ 0 < a \ll A. \]

Then \[ \text{InverseHHI} \approx n_{\text{large}} + O\left(\frac{n_{\text{small}} A}{A}\right) \quad (2). \]

The last property (2) is the most important. It shows that if the set \{x_i\} has \( n_{\text{large}} \) large values and \( n_{\text{small}} \) small values, then InverseHHI is a little bit larger than \( n_{\text{large}} \).

Similar statistics that is used in ecology is a Simpson’s diversity index[4].

3. Generalization of Inverse Herfindahl-Hirschman Index

In (1) instead of power 2 we can use arbitrary \( k > 0 \):

\[ E_k = \begin{cases} \left(\sum_{i=1}^{n} p_i^k\right)^{-\frac{1}{k}} & \text{if} \quad k \neq 1, \\ \exp\left(-\sum_{i=1}^{n} p_i \ln p_i\right) & \text{if} \quad k = 1 \end{cases} \quad (3) \]

where \( p_i = x_i / \sum_{i=1}^{n} x_i \).

The second equation in (3) for case \( k = 1 \) is chosen to have \( E_k \) continuous at \( k = 1 \).

Equations (3) define generalized exponential entropy(DE) of order \( k \) for a sample \{x_1, x_2, \ldots, x_n\} [1], that in special case \( k = 1 \) is ordinary exponential entropy (Shannon index) [5], and in special case \( k = 2 \) we get \( E_2 = \text{InverseHHI} \).

We can interpret \( p_i \) as probabilities, e.g. in case of HHI, where \( p_i \) are market shares of enterprises revenues, the \( p_i \) could be considered as probability that randomly chosen dollar of revenue belongs to enterprise i.

In frame of the probability interpretation, equations (3) define generalized exponential entropy, or distribution extents (DE) of order \( k \) for a probability distribution \{p_i\} [1], that in special case \( k = 1 \) is ordinary exponential entropy (Shannon index) [5], and in special case \( k = 2 \) we get \( E_2 = \text{InverseHHI} \), so could be named “distribution extents” of the set \{x_1, x_2, \ldots, x_n\}.

It’s important, that if distribution of revenue has density \( p_{\text{Rev}} \) then \{p_i\} are normalized quantile of empirical cumulative distribution function \( P_{\text{Rev}} \): \( p_i = x_i / \sum_{i=1}^{n} x_i \), \( x_i \approx P_{\text{Rev}}^{-1}\left(1 - \frac{i-1/2}{n}\right) \), rather than \( p_{\text{Rev}} \).
4. Properties of Distribution Extents (DE)

For an arbitrary $k$ the properties (1), (2) take form:

\[ 1 \leq E_k \leq n; \]

$E_k = 1$, when $x_1 = A, x_i = 0, i = 2, \ldots, n$ \hspace{1cm} (4a)

$E_k = n$, when $x_1 = x_2 = \ldots = x_n$ \hspace{1cm} (4b)

Let $x_i = A$, $i = 1, 2, \ldots, n_{\text{Large}}$

$x_i = a$, $i = n_{\text{Large}} + 1, \ldots, n$; where $a \ll A$

Then $E_k \approx n_{\text{Large}} \left( 1 + O \left( \frac{n_{\text{Small}}}{n_{\text{Large}}} \frac{a}{A} \right) \right)$ \hspace{1cm} (5)

If $0 < k < m$, then

\[ \frac{1}{\max(p_i)} = E_{+\infty} \leq E_m \leq E_k \leq E_0 = n \] \hspace{1cm} (6)

Property (5) shows that as in special case $k = 2$, if the set $\{x_i\}$ has $n_{\text{Large}}$ large values and $n_{\text{Small}}$ small values, then in reasonable assumptions \textit{InverseHHI} is a little bit larger than $n_{\text{Large}}$.

5. Relationship between Distribution Extents and Rényi entropy

The Rényi entropy [6, 7] of order $k$, where $k \geq 0, k \neq 1$ is defined as

\[ H_k = \frac{1}{1-k} \log(\sum p_i^k) \] \hspace{1cm} (7)

So

\[ E_k = \exp(H_k) \] \hspace{1cm} (8)

Example 1. DE for lognormal distribution.

![Figure 1. Positions of GEE $E_k$ for $k \in \{1/2, 1, 3/2, 2, \ldots, 99\}$ on reverse quantile plot of lognormal distributed $x_i, i = 1, \ldots, 1000$](image)

In this example we created sample of 1000 lognormal distributed values $x_i$, sorted them in decreasing order, plotted vs. index $i$ and marked cut-off points for values $k \in \{1/2, 1, 3/2, 2, \ldots, 99\}$. For given $k$ “large x” are left of corresponding vertical reference line.

Example 2. List of large countries (by population and GDP) for different choice of parameter $k$. 
In list of large countries by population with given value of $k$ are countries which symbols are laying above corresponding horizontals reference line on the plot Fig.1.

In list of large countries by GDP with given value of $k$ are countries which symbols are laying wright of corresponding vertical reference line on the plot Fig.2.

Example 3. Now consider distribution of population by urban agglomerations / cities, using UN data for agglomerations with population more than 750000 [8]. For each country that has at least 5 cities, we calculate $DE_k$ ("number of large cities") and corresponding population thresholds for $k=2$ and $k=99$:

![Population vs GDP by Country](image)

**Figure 2.** Positions $E_k$ for $k \in \{\frac{1}{2}, 1, 2, 4, 99\}$ on quantile plot of Population vs. GDP for major countries.
Figure 3. $D_{E_k}$ (“number of large cities”) and corresponding population thresholds for $k=2$ and $k=99$. 
We see that in both cases plots have common characteristics: China (Ch) has many large cities with smaller population in each of them, Japan (Jp) has very high concentration of population in a few huge cities with, US, Indonesia and Brasilia are somewhere in the middle.


We intensively used DE to analyze the Microsoft web advertising network of hundreds millions users, advertisers, and publishers that have hierarchical structure and are linked by page views, ad impressions, clicks, conversions and other characteristics that could be approximated as tripartite weighted graph with five types of edge weights. The main purpose of the analysis was community discovery for click fraud (collusion) detection. We can not describe the algorithm in use in details, because it makes “job” of fraudsters easier, so we provide here only general description of usage DE in the algorithms.

A. DE as a cut off parameters

DE as cut off parameters to present the network as a graph to analyze and visualize the network and to use graph theory methods for click fraud detection.

Parameter k of DE can be used as adjustment parameter in modeling. The picture below shows a fragment of the network graph with one type of edges, where threshold for the edges weights were chosen via DE with a high value of parameter k – keeping only edges connecting nodes with their “largest neighbors”.
Figure 4. A fragment of the network graph with one type of edges, where threshold for the edges weights were chosen via DE with a high value of parameter k.
1) **DE as an independent variable in predictive modeling.**

If we have to create a predictive model for countries, then we can use DE from example 3 above as predictive variables. Parameter $k$ of DE can be used as adjustment parameter in modeling in this case too. The optimal value for this parameter can be found in process of cross-validation by maximization of area under the lift curve, or another objective function.

The following example illustrates usage of DE in click fraud modeling.

<table>
<thead>
<tr>
<th>extent.NC</th>
<th>extent.F</th>
<th>trend/week</th>
<th>CV</th>
<th>main.DURL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.271</td>
<td>12.239</td>
<td>-0.0403</td>
<td>0.3451</td>
<td>cooperatefinance.com</td>
</tr>
<tr>
<td>4.546</td>
<td>6.977</td>
<td>-0.0114</td>
<td>0.6076</td>
<td>topairplay.com</td>
</tr>
<tr>
<td>2.349</td>
<td>4.408</td>
<td>0.0113</td>
<td>1.5730</td>
<td>goodsearch.com</td>
</tr>
<tr>
<td>3.604</td>
<td>9.758</td>
<td>0.0103</td>
<td>0.7084</td>
<td>trainingtopic.com</td>
</tr>
<tr>
<td>3.082</td>
<td>5.096</td>
<td>-0.1361</td>
<td>0.3491</td>
<td>holidaysf.com</td>
</tr>
<tr>
<td>3.460</td>
<td>5.249</td>
<td>-0.0144</td>
<td>0.6814</td>
<td>companypetroleum.com</td>
</tr>
<tr>
<td>2.987</td>
<td>4.291</td>
<td>-0.1623</td>
<td>0.5622</td>
<td>idgeuk.com</td>
</tr>
<tr>
<td>13.156</td>
<td>13.989</td>
<td>0.0044</td>
<td>0.0204</td>
<td>facebook.com</td>
</tr>
<tr>
<td>1.112</td>
<td>6.269</td>
<td>0.0001</td>
<td>1.1995</td>
<td>autofinancecompanies-au.com</td>
</tr>
</tbody>
</table>

**Figure 5.** A fragment of a table with four predictive independent variables describing time series for a set of domains.

Variables *extent.NC* and *extent.F* are time independent DE giving number of large values in time series of some other time dependent variables NC and F for a specified time interval. The plot of a time series, where variable NC corresponds to size of bubbles, and variable F corresponds to the vertical axis, looks as following:
2) DE as a criterion for optimization

The third way of usage DE as a criterion for optimization of some parameters of models, could be demonstrated with following example.

When we do clustering of \( n \) objects using k-means algorithm, we often get one huge cluster and many very small clusters: \( x_1 \gg x_2, \ldots, x_n \), where \( x_i \) is a number of objects in cluster \( i \). That does the result of clustering not very useful. Changing parameters of clustering, e.g. adding other variables or changing similarity/distance function, we get other variants of clustering. A good criterion to choose variant of clustering could be maximizing value of DE for the sets \( \{x_i\} \) of numbers of objects in clusters (characterizing number of “large clusters”)

7. Acknowledgment

Thanks to Gang Wu, Jigar Mody, Jingying Zhang, Dennis Minium, Qiang Li and Anil Biradar for fruitful discussions.

8. References

