Empirical Data Analytics

Plamen Angelov, Xiaowei Gu, Dmitry Kangin

School of Computing and Communications

Lancaster University

Lancaster, LA1 4WA, UK

e-mail: {p.angelov, x.gu3, d.kangin}@lancaster.ac.uk

Abstract:

In this paper, we propose an approach to data analysis which is based entirely on the empirical observations of discrete data samples and the relative proximity of these points in the data space. At the core of the proposed new approach is the *typicality* - an empirically derived quantity which resembles probability. This non-parametric measure is a normalised form of the *square centrality* (*centrality* is a measure of closeness used in graph theory). It is also closely linked to the *cumulative proximity* and *eccentricity* (a measure of the tail of the distributions that is very useful for anomaly detection and analysis of extreme values). In this paper, we introduce and study two types of *typicality*, namely local and global versions. The *local typicality* resembles the well-known probability density function (*pdf*), probability mass function and fuzzy set membership but differs from all of them. The *global typicality*, on the other hand, resembles well-known histograms but also differs from them. A distinctive feature of the proposed new approach, Empirical Data Analytics (EDA), is that it is not limited by restrictive impractical *prior* assumptions about the data generation model as the traditional probability theory and statistical learning approaches are. Moreover, it does not require an explicit and binary assumption of either randomness or determinism of the empirically observed data, their independence or even their number (it can be as low as couple of data samples). The *typicality* is considered as a fundamental quantity in the pattern analysis, which is derived directly from data and is stated in a discrete form in a contrast to the traditional approach where a continuous *pdf* is assumed *a priori* and estimated from data afterwards. The *typicality* introduced in this paper is free from the paradoxes of the *pdf*. *Typicality* is objectivist while the fuzzy sets and the belief-based branch of the probability theory are subjectivist. The *local typicality* is expressed in a closed analytical form and can be calculated recursively; thus, computationally very efficiently. The other non-parametric ensemble properties of the data introduced and studied in this paper, namely, the *square centrality*, *cumulative proximity* and *eccentricity* can also be updated recursively for various types of distance metrics.

Finally, a new type of classifier called naïve Typicality-based EDA class is introduced which is based on the newly introduced *global typicality.* This is only one of the wide range of possible applications of EDA including, but not limited for anomaly detection, clustering, classification, control, prediction, control, rare events analysis, etc. which will be the subject of further research.

Index terms — data mining and analytics, probability, statistics, pattern recognition, machine learning, local and global typicality, eccentricity, centrality.

I. Introduction

Data analysis can be described as a process which applies statistical and/or formal techniques to describe, illustrate and evaluate data. It became a hot topic recently in many different areas such as biology, econometrics, epidemiology, social science, social media, cyber-security, and so on. Nowadays, new scientific areas are becoming data-centric (if previously data rich were mostly the engineering, natural sciences and, to some extent, economics, now biomedical, social and other sciences are also increasingly becoming data-centric). There is a growing demand for alternative new concepts for data analysis that are centred at the actual data collected from the real world rather than at theoretical *prior* assumptions which need then to be confronted for verification with the experimental data as is the case with the traditional statistical approach [1]-[4]. The traditional probability theory and statistics assume the actual data to be realizations of imaginary random variables and further assume the *prior* distributions of these variables.

The general problem of probability theory was defined by Kolmogorov as follows: "*Given a cdf F(x), describe outcomes of random experiments for a given theoretical model.*" [5]. Vapnik and Izmailov define the general problem of statistics as follows: "*Given iid observations of outcomes of the same random experiments, estimate the statistical model that defines these observations*" [6]. Both, traditional probability theory and statistics have strong and often impractical requirements and assumptions ("*Given a cdf …*"; "*Given iid, …same random experiments*", etc.). They also assume a random nature for the variables which is indeed the case for some problems, such as gambling, independent experts, etc. However, real processes of interest (such as climate, economic, social, mechanical, electronic, biological, etc.) are complex and not always display a clear (deterministic or stochastic) nature. Both, the traditional probability theory and statistics have strong and often impractical requirements and assumptions ("*Given a cdf …*"; "*Given iid, …same random experiments*", etc.). They also assume random nature of the variables which is indeed the case for problems, such as gambling, games, independent experts, etc.

However, real processes of interest (such as climate, economic, social, mechanical, electronic, biological, etc.) are complex and not always with a clear nature (deterministic or stochastic). A more recent alternative is to approximate the distributions using non-parametric, data-centered functions, such as particle filters [7], entropy-based information-theoretic learning [8], etc. On the other hand, partially trying to address the same problems, in 1965 L. Zadeh introduced fuzzy sets theory [9] which completely departed from objective observations and moved (similarly to the belief-based theory [10] introduced a bit later) to the subjectivist definition of the uncertainty. A later strand of fuzzy set theory (data driven approach) developed mainly in 1990s attempted to define the membership functions based on experimental data which stands in between probabilistic and fuzzy representations [11, 12], however, this approach requires assuming the type of membership function.

Within the Empirical Data Analytics (EDA), we define the main problem as follows: "*Given observations of outcomes of real processes/experiments alone, estimate the ensemble properties of the data, such as cumulative proximity, eccentricity, density and typicality of the data. Furthermore, estimate these for any feasible outcome*".

In this paper, we introduce novel non-parametric estimators of ensemble statistical properties of the data derived entirely from the experimental discrete observations. These include the *square centrality*, *eccentricity* (ξ), *standardized eccentricity* (ε) as well as the *local* ***typicality***, (τ). The newly proposed non-parametric estimators are defined for various distance metrics. Furthermore, in this paper we introduce the *global* ***typicality*** (τ^G) which is similar (but different) to the histograms. The *typicality* looks very similar to the well-known *pdf,* it sums up to *1* and is always positive; however, it is discrete and is always less than *1* while the *pdf* can paradoxically be greater than *1*. Additionally, the *typicality* is only defined for feasible values of the independent variable while *pdf* can, paradoxically, be positive even for infeasible values, e.g. negative height, distance, weight, absolute temperature, etc.

The remainder of the paper is organized as follows. In section II, we introduce the basic elements of the Empirical Data Analysis (EDA). The concepts of the *local* and *global typicality* as well as estimation for hypothetical new points are described in section III. The properties of the EDA quantities are discussed in section IV. In section V an example of the newly proposed naïve EDA classifier is described and, finally, section VI concludes the paper outlining the future development.

II. Basic Elements of the Empirical Data Analysis

In this section, the basics of the Empirical Data Analytics approach are introduced, including:

a) *cumulative proximity*, q and its inverse, called *square centrality*, S ;

b) *eccentricity*, ξ and *standardised eccentricity*, ε;

c) *density*, D;

d*) local* and *global* ***typicality,*** τ, τ^G.

The *global* ***typicality,*** τ^G addresses the global properties of the data and will be introduced in the next section.

First, let us define the data space, M with a distance $d(\cdot,\cdot)$. Let us consider a finite multiset of data points $\aleph_k = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k\}, \quad \boldsymbol{x}_i \in \aleph_k, \quad 1 \le i \le k,$ where the index k denotes the amount of data samples/points/vectors. p is the dimensionality of the data.

A. Centrality

First, we will recall from the graph and network theory the so called measure of *centrality*, C [13,14]. Indeed, for every point $\boldsymbol{x}_i \in \aleph_k$ one may need to quantify how *close* or *similar* this point is to ***all*** other data points from $\aleph_k$. In graph (networks) theory a measure of *centrality* [13] is defined as the inverse of the so called *farness* which itself is a sum of distances from a point to ***all*** other points:

$$C_k(\boldsymbol{x}_i) = \frac{1}{\sum_{j=1}^{k} d(\boldsymbol{x}_i, \boldsymbol{x}_j)} = \left(\sum_{j=1}^{k} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^{-1}; \quad k > 1; \quad \boldsymbol{x}_i \in \aleph_k; \quad 1 \le i \le k; \qquad \forall \boldsymbol{x}_j \left| \boldsymbol{x}_j \ne \boldsymbol{x}_i \right. \tag{1}$$

Notice that in this and further notations, k denotes the amount of data samples at the moment of the analysis. For the offline cases this is usually denoted by N to represent the total amount of data available and is not changing throughout the analysis while in online cases the data is a stream rather than a set and k may or may be associated with the order of data samples which usually but nt necessarily always is the time instant.

B. Cumulative proximity

In our earlier works [15-16] we defined a measure called *cumulative proximity*, $q_k(\boldsymbol{x}_i) \quad \boldsymbol{x}_i \in \aleph_k$ which can be seen as a square form of the *farness*, as follows:

$$q_k(\boldsymbol{x}_i) = \sum_{j=1}^{k} d^2(\boldsymbol{x}_i, \boldsymbol{x}); \quad k = |\aleph_k| > 1; \tag{2}$$

The *cumulative proximity*, plays an important role in the definition of the other association measure derived empirically from the observed data without making any *prior* assumptions about their generation model, namely the *typicality and eccentricity* as we will see later.

C. Square centrality

It is convenient to consider the *square centrality* defined as inverse of the *cumulative proximity*:

$$S_k\left(\boldsymbol{x}_i\right)=\frac{1}{q_k\left(\boldsymbol{x}_i\right)}=q_k^{-1}\left(\boldsymbol{x}_i\right);\quad k=\left|\aleph_k\right|>1 \tag{3}$$

D. Eccentricity

The eccentricity, $\xi_k\left(\boldsymbol{x}\right)$ is defined within EDA as a *normalized cumulative proximity* [15,16]. It is a very important measure of the ensemble property related to the tail of the distribution and is empirically derived from the observed data only without making any *prior* assumptions about their generation model. It plays an important role in anomaly detection [16], analysis of rare events as well as for the estimation of the *typicality* as it will be detailed further. The eccentricity of a particular data sample $\boldsymbol{x}_i$ ($\boldsymbol{x}_i\in\aleph_k$) is calculated as follows:

$$\xi_k\left(\boldsymbol{x}_i\right)=\frac{2q_k\left(\boldsymbol{x}_i\right)}{\sum_{j=1}^{k}q_k\left(\boldsymbol{x}_j\right)};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1; \tag{4}$$

where, the coefficient *2* is included to compensate distance duplication in the denominator which, at the same time, leads to the following bounds for the eccentricity value:

$$0\le\xi\left(\boldsymbol{x}_i\right)\le 1;\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{5}$$

The denominator must not be zero; from the metric definition it follows that at least two points within the dataset $\aleph_k$ need to be distinctive. Another property of the *eccentricity* is that it sums over $\aleph$ to *2*:

$$\sum_{i=1}^{k}\xi\left(\boldsymbol{x}_i\right)=2;\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{6}$$

In EDA we further introduced [15,16] the *standardized eccentricity,* $\varepsilon_k\left(\boldsymbol{x}\right)$ which normalises the cumulative proximity by half of the average *cumulative proximity*, where the coefficient *2* is used for the same reasons as mentioned above:

$$\varepsilon_k\left(\boldsymbol{x}_i\right)=k\xi_k\left(\boldsymbol{x}_i\right)=\frac{2q_k\left(\boldsymbol{x}_i\right)}{\frac{1}{k}\sum_{j=1}^{k}q_k\left(\boldsymbol{x}_j\right)};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{7}$$

Standardised eccentricity is more convenient to use in the well-known Chebyshev inequality which within TEDA has a more elegant form.

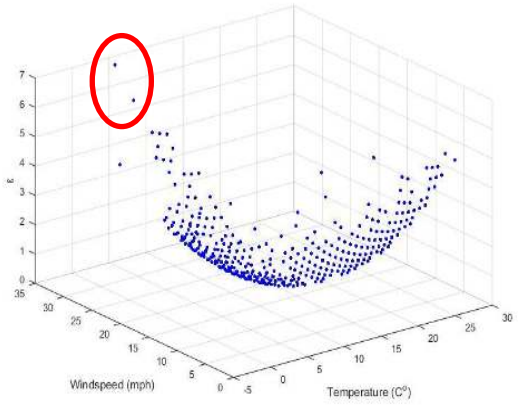


Fig.1 *Standardized eccentricity*, ε for real climate data (temperature and wind speed) measured in Manchester, UK for the period 2010-2015 [19].The data with higher standardized eccentricity, ε are marked with red ellipsoids; these are days with stronger wind (stormy days). As it will be described in the next subsection, values of ε between *5* and *10* correspond to so called $\left[2\sigma;\ 3\sigma\right]$ interval from the mean.

E. Density

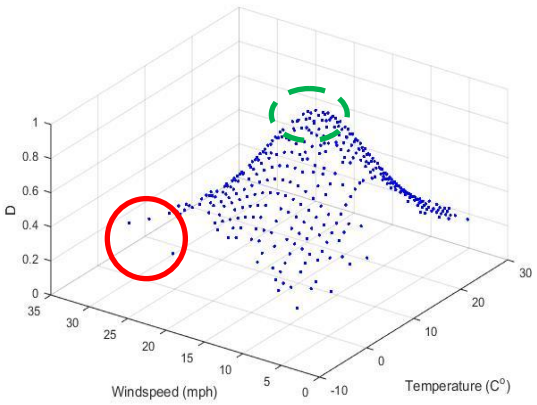


Fig. 2 The density for the same real climate data as shown in Figure 1 [19]. The data with lower *density D* are marked with red ellipsoids. The data with high density are marked with a green dashed line ellipsoid.

The *density* within the data space plays an important role in data analysis [15-17]. Within EDA it can be defined as the inverse of the *standardised eccentricity*:

$$D_k\left(\boldsymbol{x}_i\right)=\varepsilon_k^{-1}(\boldsymbol{x}_i);\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{8}$$

F. Chebyshev inequality

The Chebyshev inequality [18] is well known in the traditional probability theory and statistics. It describes the probability that certain data sample, x is more than $n\sigma$ distance away from the mean where σ denotes the standard deviation [1-3] and can be formulated as follows [18] if use Euclidean type of distance:

$$P\left(\left\|\boldsymbol{\mu}_k-\boldsymbol{x}_i\right\|^2\le n^2\sigma_k^2\right)\ge 1-\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{9}$$

Respectively, the probability the point x_i to be an outlier is given by:

$$P\left(\left\|\boldsymbol{\mu}_k-\boldsymbol{x}_i\right\|^2> n^2\sigma_k^2\right)<\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{10}$$

It can be proven that ***exactly*** the same result can be provided within EDA through the *standardized eccentricity* [16] for the Euclidean distance:

$$P\left(\varepsilon_k(\boldsymbol{x}_i)\le n^2+1\right)\ge 1-\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{11}$$

and, respectively,

$$P\left(\varepsilon_k(\boldsymbol{x}_i)> n^2+1\right)<\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{12}$$

Similarly, the Chebyshev inequality in the form of *density* are as follows.

$$P\left(D_k(\boldsymbol{x}_i)\ge\frac{1}{n^2+1}\right)\ge 1-\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{13}$$

$$P\left(D_k(\boldsymbol{x}_i)<\frac{1}{n^2+1}\right)<\frac{1}{n^2};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{14}$$

The attractiveness of the equations (11)-(14) in comparison with the equations (9)-(10) is that **no *prior* assumptions** are required within EDA on the nature of the data (random or deterministic), the generation model, the amount of data and their independence. In addition, the result is more elegant and a similar result can be derived for Mahalanobis and other type distance metrics [16].

G. Recursive calculations

One important aspect of the proposed EDA approach is that the operators defined within it can be calculated recursively for various Hilbert space metrics, e.g. Euclidean, Mahalanobis, even value Minkowski as well as for cosine similarity, which makes EDA suitable for live data streams processing. Consider the time instance the data point x_k arrives, it is easy to see that:

$$q_k\left(\boldsymbol{x}_i\right)=q_{k-1}\left(\boldsymbol{x}_i\right)+d^2\left(\boldsymbol{x}_i,\boldsymbol{x}_k\right);\quad k>1;\quad \boldsymbol{x}_i\in\aleph_k;\quad 1\le i\le k-1 \tag{15}$$

and respectively

$$S_k\left(\boldsymbol{x}_i\right)=\left(S_{k-1}^{-1}\left(\boldsymbol{x}_i\right)+d^2\left(\boldsymbol{x}_i,\boldsymbol{x}_k\right)\right)^{-1};\quad k>1;\quad \boldsymbol{x}_i\in\aleph_k;\quad 1\le i\le k-1 \tag{16}$$

1) Euclidean type metric

For example, for the Euclidean type distances we have [15-17]:

$$\begin{aligned} q_k\left(\boldsymbol{x}_i\right)&=k\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\mu}_k\right)\\ &=k\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+\sigma_k^2\right);\quad k>1;\quad \boldsymbol{x}_i\in\aleph_k;\quad 1\le i\le k \end{aligned} \tag{17}$$

where $\sigma_k^2=X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\mu}_k$. And, respectively:

$$\begin{aligned} S_k\left(x_i\right)&=\frac{1}{k\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\mu}_k\right)}\\ &=\frac{1}{k}\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\mu}_k\right)^{-1};\quad k>1;\quad \boldsymbol{x}_i\in\aleph_k;\quad 1\le i\le k \end{aligned} \tag{18}$$

$$\boldsymbol{\mu}_k=\frac{k-1}{k}\boldsymbol{\mu}_{k-1}+\frac{1}{k}\boldsymbol{x}_k;\quad \boldsymbol{\mu}_1=\boldsymbol{x}_1 \tag{19}$$

$$X_k=\frac{k-1}{k}X_{k-1}+\frac{1}{k}\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k;\quad X_1=\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 \tag{20}$$

The total *square centrality* of the whole dataset and the total *cumulative proximity*, respectively, can also be updated recursively for each new data point, $\boldsymbol{x}_k$:

$$\sum_{i=1}^{k}q_k\left(\boldsymbol{x}_i\right)=\sum_{i=1}^{k-1}q_{k-1}\left(\boldsymbol{x}_i\right)+2q_k\left(\boldsymbol{x}_k\right)=2k\left(X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\mu}_k\right);\quad q_1\left(\boldsymbol{x}_1\right)=0 \tag{21}$$

Similar equations can be written for the case of $\boldsymbol{U}$.

2) Mahalanobis distance

For the case of Mahalanobis type distance recursive update is also possible [15-17]:

$$q_k\left(\boldsymbol{x}_i\right)=k\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \tag{22}$$

and, respectively:

$$S_k\left(\boldsymbol{x}_i\right)=\frac{1}{k}\left(\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)+X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right)^{-1} \tag{23}$$

where the covariance matrix $\boldsymbol{\Sigma}_k$ is defined as follows:

$$\boldsymbol{\Sigma}_k=\frac{1}{k-1}\sum_{i=1}^{k}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}} \tag{24}$$

which itself can be updated recursively [1]-[3]. However, the average scalar product X_k is defined differently from the case when Euclidean distance is used, namely, $X_k=\frac{1}{k}\sum_{i=1}^{k}x_i^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}x_i$. Thanks to the covariance matrix symmetricity [20], there is:

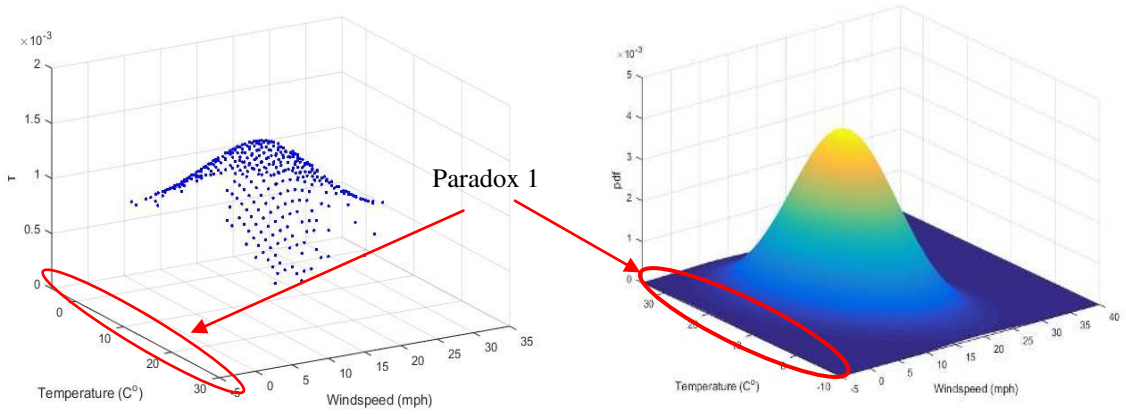
$$\sum_{i=1}^{k}q_k\left(\boldsymbol{x}_i\right)=\sum_{i=1}^{k-1}q_{k-1}\left(\boldsymbol{x}_i\right)+2q_k\left(\boldsymbol{x}_k\right)=2k^2\left(X_k-\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right)=2k^2p;\quad q_1\left(\boldsymbol{x}_1\right)=0 \tag{25}$$

III. Typicality and Estimation for Hypothetical Data Points

A. Local typicality

In EDA we define the *local typicality*, τ of a data point as *normalized square centrality* or, which is the same, *normalized inverse cumulative proximity*:

$$\tau_k\left(\boldsymbol{x}_i\right)=\frac{S_k\left(\boldsymbol{x}_i\right)}{\sum_{j=1}^{k}S_k\left(\boldsymbol{x}_j\right)};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{26}$$

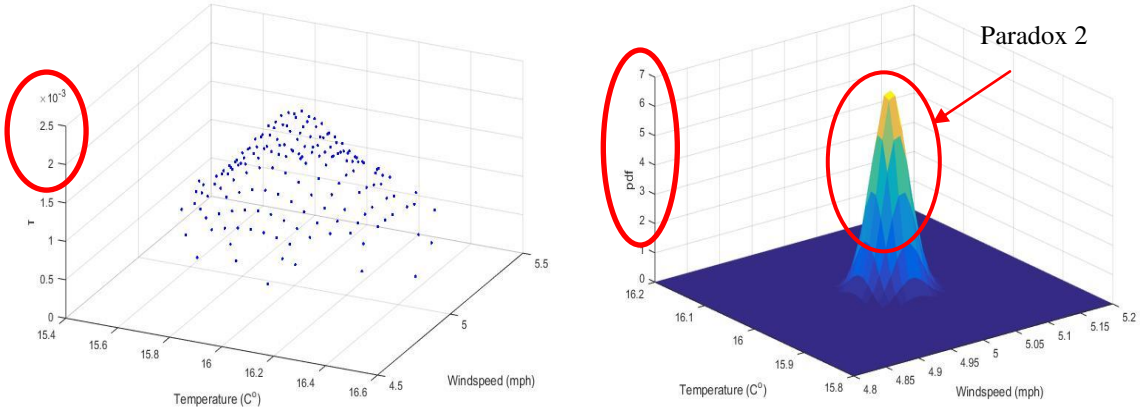


(a) *Local typicality*, τ (b) *pdf*

Fig. 3 A comparison of the *local typicality* distribution, $\tau(x)$ and the traditional *pdf* based on the same real climate data [19] as in Fig. 1 and 2. The red oval indicates infeasible values of negative wind speed which have *pdf>0* according to the *pdf*. τ is only defined for feasible values (in this case only for positive values of the wind speed).

It is very interesting to stress that the *local typicality* resembles the well-known *pdf*, but is free from paradoxes that can be the case with the *pdf*. Figure 3 visualizes for the same real data [19] the paradox 1 that a Gaussian type *pdf* may have non-zero values for infeasible values (e.g. negative wind speed).

Fig. 4 further demonstrates another paradox (paradox 2) that *pdf* values can be >*1*.



(a) *Local typicality*, τ (b) *pdf*

Fig. 4 Examples of the paradox 2 regarding the temperature and wind speed. The red ellipsoid indicates the area exhibiting infeasible values of *pdf* that *pdf* >1.

B. Global typicality

It is well known that the traditional *pdf* expressed as Gaussian or Cauchy function are not perfect to cover real distributions. It is also a well-known technique to use mixtures of distributions [21]. Within EDA it is possible to use a mixture of *local typicality* distributions derived after clustering the data [22], however, it is also possible to derive a *global typicality*, τ^G from data directly without clustering or any other pre-processing or *prior* assumptions. In this paper we define the *global typicality*, τ^G as a *weighted normalised square centrality* where the weights are the frequency of occurence of a particular data sample, *f*.

Let us first introduce couple of definitions. For each dataset $\aleph_k$, one can construct a corresponding unique data points set as $\boldsymbol{U}_k = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_i, ..., \boldsymbol{u}_l\}$ ($\boldsymbol{u}_i \in \aleph_k$; where *l* is the number of the unique data samples) and the corresponding numbers of times of occurrence by $\boldsymbol{F}_k = \{f_1, f_2, ..., f_i, ..., f_l\}$. We can also view f_i as a frequency and optionally divide by *k*. if we prefer values that are ≤ 1. Then the *global typicality*, τ^G can be defined within EDA on the domain of $\boldsymbol{U}_k$ as follows:

$$\tau_k^G\left(\boldsymbol{u}_i\right) = \frac{f_i S_k^u\left(\boldsymbol{u}_i\right)}{\sum_{j=1}^{l} f_j S_k^u\left(\boldsymbol{u}_j\right)}; \quad \sum_{j=1}^{l} f_j S_k^u\left(\boldsymbol{u}_j\right) > 0; \quad l > 1; \quad 1 \leq i \leq l \tag{27}$$

where $S_k^u\left(\boldsymbol{u}_i\right) > 0$ denotes the *square centrality* at a particular unique data sample $\boldsymbol{u}_i$ to all other unique data samples of the data space.

Some illustrative examples of applying the *global typicality* are provided further. It is interesting to note that for small values of *k*, the *global typicality* is ***exactly the same*** as the frequentistic form of probability (see Fig. 3 and 4), and with large *k*, the *local typicality* approximates the well-known Gaussian and Cauchy type *pdf*.

Let us start with a trivial primer and consider a small data set $\aleph = \{2;6;2;2;6\}$. It is also straightforward to determine $\boldsymbol{U}_5 = \{2;6\}; l = 2;\ k = 5; l < k;\ \boldsymbol{F}_5 = \{3;2\}$. Let us consider Euclidean type distance. The *global typicality* is easy to calculate for this case using equation (3) for $S = \{1/16;1/16\}$ and equation (27) for τ^G but taken only in regards to the unique data, $\boldsymbol{U}$ instead of all data, $\aleph$. The final result, $\tau^G\left(x=2\right) = 3/5 = 0.6$ and $\tau^G\left(x=6\right) = 2/5 = 0.4$ is exactly the same as the frequentistic probability expressed as a ratio of specific outcome to the total number of outcomes [1-4], see Fig. 5.

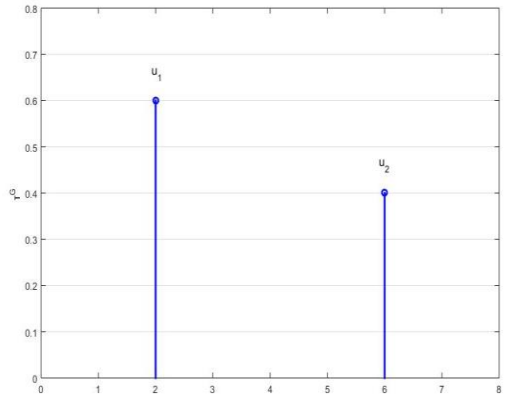


Fig. 5 *Global typicality*, τ^G directly calculated from the data for the trivial primer

Another example where the *global typicality* has ***exactly*** the same values as the frequentistic probability is if consider fair games which are clearly random. It has to be stressed that the encoding of the values is important. For example, if consider tossing a coin it is easy to encode the two possible outcomes by any number, but if consider a dice which has 6 possible outcomes these usually have an integer number of dots on them (*1,2,3,4,5,* and *6)*. Because within EDA the distance between different outcomes is also taken into account as well as the frequency, *f* the possible outcomes have to be first encoded in such a way that the distance between each pair of outcomes is exactly the same. For example, for *1* we can use $\{1;0;0;0;0;0\}$, for *2* we can use $\{0;1;0;0;0;0\}$, etc. Then if we have throwing a dice *100* times (or *100* people throwing a dice once) the possible outcomes may be as tabulated in Table 1. It is easy to check that $S=\{0.1;0.1;0.1;0.1;0.1;0.1\}$ and $F=\{17;14;15;15;21;18\}$.

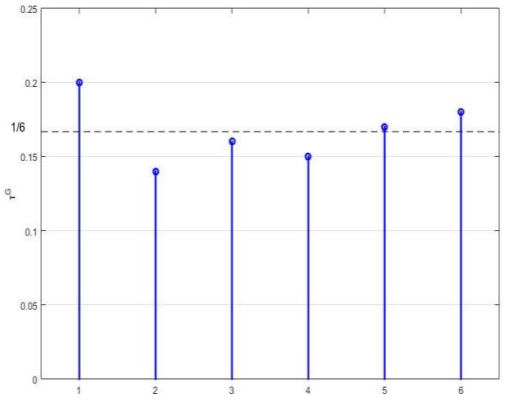


Fig. 6 A simple illustrative example of a fair game of throwing dices (random data)

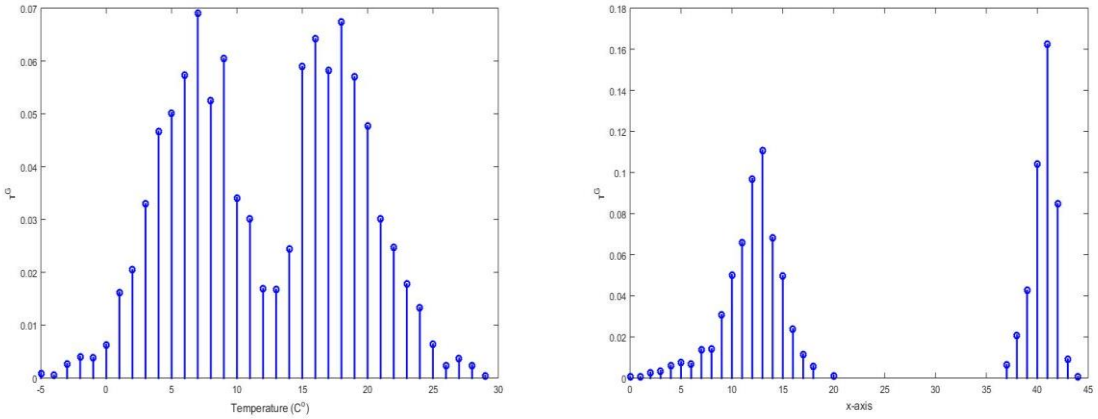
Table 1. Throwing dices (a fair game, random data)

Possible outcome	**Mapping**	**Frequency of occurrence**
1	$[1,0,0,0,0,0]$	20
2	$[0,1,0,0,0,0]$	14
3	$[0,0,1,0,0,0]$	16
4	$[0,0,0,1,0,0]$	15
5	$[0,0,0,0,1,0]$	17
6	$[0,0,0,0,0,1]$	18
Total		100

The *global typicality* of each possible outcome is depicted in Fig. 6. It is easy to see that the values of τ^G fluctuate around the value of 1/6 and they would have all have value of 1/6 if the frequencies of occurrence of each value were the same. The more times we throw the dice, the closer the values of the τ^G will be to the value of 1/6 ***exactly*** the same as the frequentistic probability. In this case, obviously the traditional *pdf* [1-4] will be misleading although paradoxically the infamous *iid* conditions are satisfied.

So far, we have seen two simple examples where the *global typicality* has ***exactly*** the same values as the frequentistic probability. In general, as it is well known, the majority of real processes of interest such as climate, economic, social, biological, technical etc. are not clearly random or deterministic. Even more difficult is to pre-determine the underlying distribution(s) from which samples are drawn. Therefore, a very attractive approach is to study the data pattern as it is observed and to develop techniques that do not require the user to pre-determine the randomness or determinism, the type of the distributions as well as other parameters. The aim of EDA is precisely to offer such an approach. In the next example, we will consider again the same real climate data [19] and will also consider another real data set taken from wearable wrist–worn sensors [23] (partially). It is very interesting and unique property of the *global typicality*, τ^G (Fig. 7 and 8) that for $l << k$ (when there are many different data points with the same value) different modes of the distribution starts to appear automatically and there is no need to pre-determine them or to apply clustering or optimization to identify them. It has to be stressed that although the result resembles histograms it is principally different. The values on the vertical axis are real values and the mutual proximity and centrality of the data is taken into account unlike the case of histograms. For example, if multiple (say, *10*) times the same value is added to the data set it will not have the

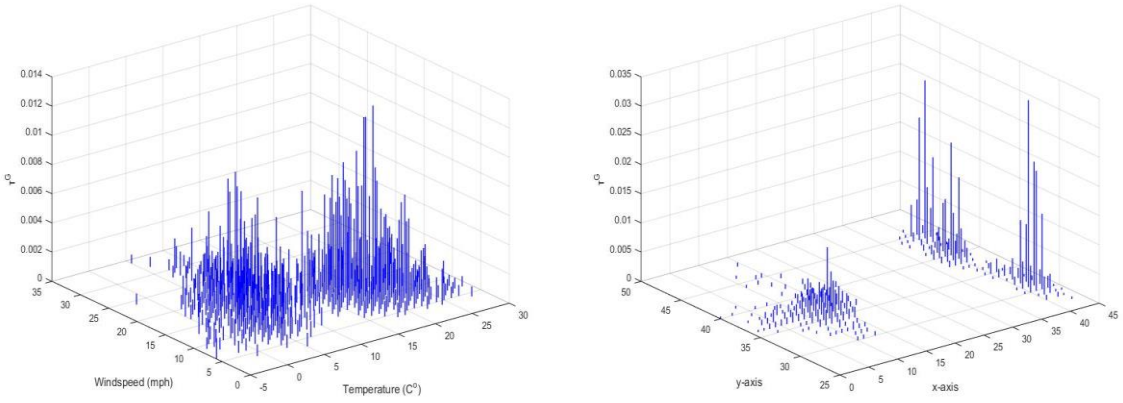
same effect if it is far from the mean (e.g. temperature 26^oC) or close to the mean (e.g. 14^oC). In the case of a traditional histogram the effect of adding multiple new samples will be the same regardless of the position. Moreover, the *global typicality* is describes by a closed analytical form equation (27).



(a) Climate dataset (temperature) (b) Wearable sensors dataset (*x*-axis)

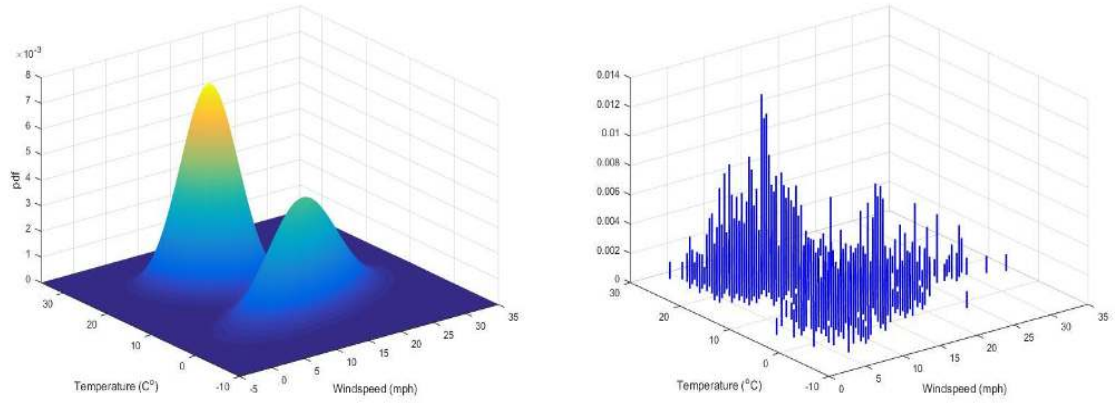
Fig. 7 *2D global typicality*, τ^G (left plot: temperature [19]; right plot: horizontal acceleration for the wearable devices data [23]).

Notice that the modes (*2* on the left plot of Fig. 7 and 5 on the right plot) appear automatically and are not pre-defined). *3D* examples (τ^G vs *2* variables) for the same data are depicted in Fig. 8.



(a) Climate dataset [19] (b) Wearable sensors dataset [23]

Fig. 8 *3D*-examples of the *global typicality* of two datasets described above



(a) Gaussian mixture *pdf* (b) Histogram

Fig. 9 An visual comparison for the same real dataset [19] as the one used in Fig. 1-4, 7a and 8a. (The left hand plot depicts the Gaussian mixture *pdf* derived by clustering; the right plot – the histogram)

C. Estimating the global typicality of hypothetical points

The mechanism described so far does represent the *global typicality* of data that have been facts (took place and is available). Often of interest is to estimate the *global typicality* (similar to the probability, likelihood) of hypothetical data points, $\boldsymbol{x}_{k+1}$. The approach we take within EDA is to consider such points (even if they are multiple) one by one to avoid accumulation of errors. In general, there are two options:

i) either the new datum $\boldsymbol{x}_{k+1}$ belongs to the corresponding unique values set $\boldsymbol{U}_l$ and one can just increment the number of occurrence within $\boldsymbol{F}_l$, or

ii) the datum does not belong to $\boldsymbol{U}_l$. In the latter case, one should append the unique data samples set from $\boldsymbol{U}_l = \{\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_l\}$ to $\boldsymbol{U}_{l+1} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_l, \boldsymbol{u}_{l+1}\}$, $\boldsymbol{u}_{l+1} = \boldsymbol{x}_{k+1}$.

After that, for Euclidean and Mahalanobis distances, one needs to update the mean value and the (co-)variance taking into account this new datum. For example, for the Euclidean case, we can recursive calculate the squared centrality of the new data point/sample as $\boldsymbol{u}_{l+1}$ given $|\boldsymbol{U}_{l+1}| = l+1$:

$$\boldsymbol{\upsilon}_{l+1} = \frac{l}{l+1}\boldsymbol{\upsilon}_l + \frac{1}{l+1}\boldsymbol{u}_{l+1}; \quad \boldsymbol{\upsilon}_1 = \boldsymbol{u}_1 \tag{28}$$

$$U_{l+1} = \frac{l}{l+1}U_{l+1} + \frac{1}{l+1}\boldsymbol{u}_{l+1}^{\mathrm{T}}\boldsymbol{u}_{l+1}; \quad U_1 = \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{u}_1 \tag{29}$$

where $\boldsymbol{\upsilon}_l$ denotes the mean value for the l unique locations in the data space;

U_l denotes the scalar product of the l unique locations in the data space.

$$S_k^u\left(\boldsymbol{u}_{l+1}\right)=\frac{1}{\left(l+1\right)}\left(\left(\boldsymbol{u}_{l+1}-\boldsymbol{\upsilon}_{l+1}\right)^{\mathrm{T}}\left(\boldsymbol{u}_{l+1}-\boldsymbol{\upsilon}_{l+1}\right)+U_{l+1}-\boldsymbol{\upsilon}_{l+1}^{\mathrm{T}}\boldsymbol{\upsilon}_{l+1}\right)^{-1} \tag{30}$$

Further, we estimate the frequency of the hypothetical data sample based on the frequencies of the nearest actual data samples. If the hypothetical data point is surrounded by actual data points from both sides per dimension (interpolation case) then we estimate the frequency as the average of the frequencies of the neighbouring data points:

$$f_{l+1}=\frac{1}{p}\sum_{i=1}^{p}\frac{\left(u_{l+1,i}-u_{L,i}\right)f_{R,i}+\left(u_{R,i}-u_{l+1,i}\right)f_{L,i}}{\left(u_{R,i}-u_{L,i}\right)} \tag{31}$$

where $u_{R,i}$ and $u_{L,i}$ are the i^{th} dimensional values of two existing unique locations that are the nearest to $\boldsymbol{u}_{l+1}$ in the i^{th} dimension and satisfy $u_{L,i}<u_{l+1,i}<u_{R,i}$.

In case, if the estimation is for a hypothetical data point that is outside of the range of the actual data (extrapolation) the frequency is set to *1*:

$$f_{l+1}=1 \tag{32}$$

Finally, the *global typicality* is estimated using (27) in regards to $\boldsymbol{u}_{l+1}$:

$$\tau_k^G\left(\boldsymbol{u}_{l+1}\right)=\frac{f_{l+1}S_k^u\left(\boldsymbol{u}_{l+1}\right)}{\sum_{j=1}^{l}f_jS_k^u\left(\boldsymbol{u}_j\right)} \tag{33}$$

x_1 x_2

To illustrate the estimation of the *global typicality* of hypothetical new data points, let us start again with a trivial primer and work it out step by step. Let us consider a trivial dataset that consists of a single variable, x with *2* experimentally observed values: $x_1=2$ and $x_2=6$. It is easy to find that:

$$k=\{1,2\};\ d_{12}=d_{21}=4;\ p_1=p_2=16;\ S_1=S_2=1/16;\ f_1=f_2=1;\ t_1=t_2=1/2$$

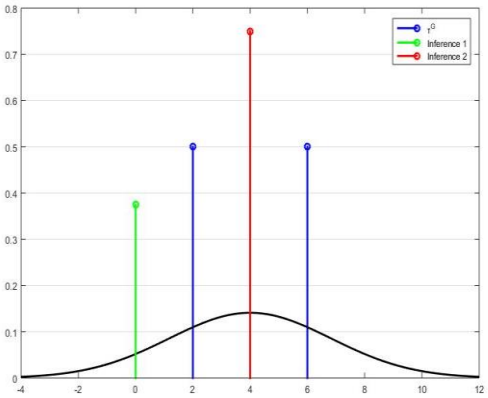
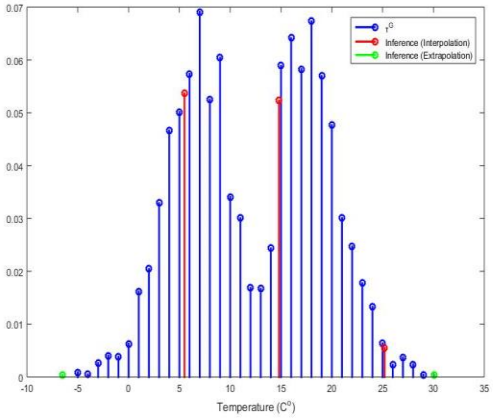
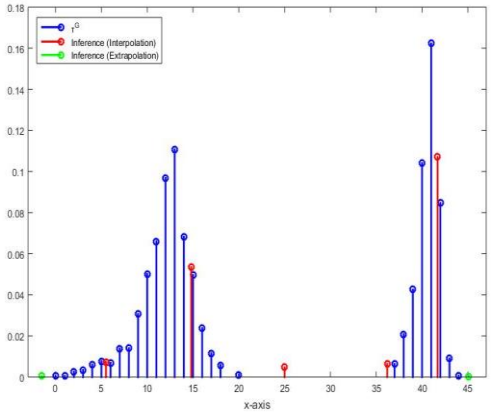


Fig. 10 Illustrative example of estimation of the *global typicality* for hypothetical data points (in red and green); a comparison with the traditional Gaussian *pdf* for our trivial primer.

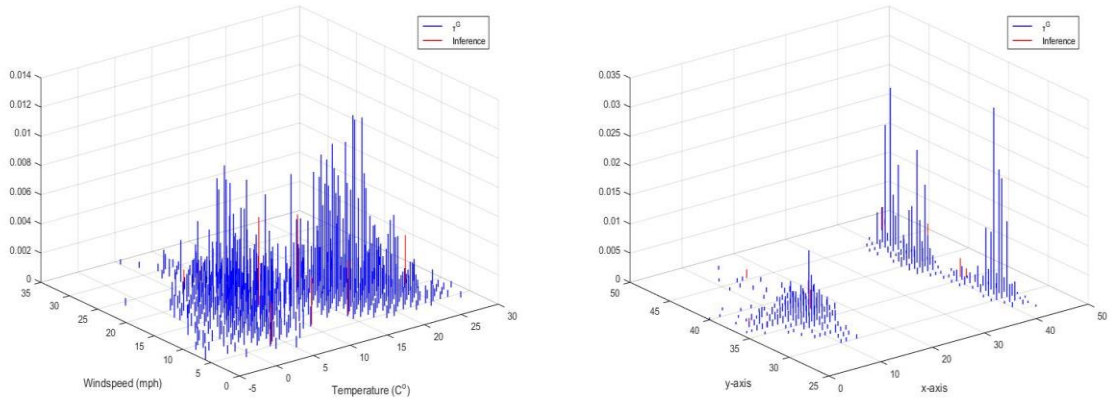


(a) Climate dataset (Temperature)



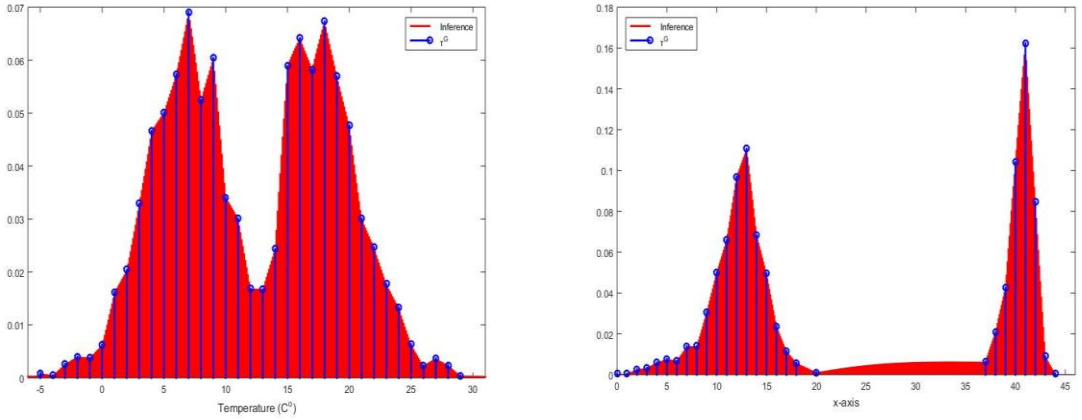
(b) Wearable sensors dataset (x-axis)

Fig.11 Examples of estimation (interpolations and extrapolation) of *global typicality* for the climate [19] and wearable sensors datasets [23] and for hypothetical data points (the blue points represents the actual points; the green points represent interpolation; red points – extrapolation)

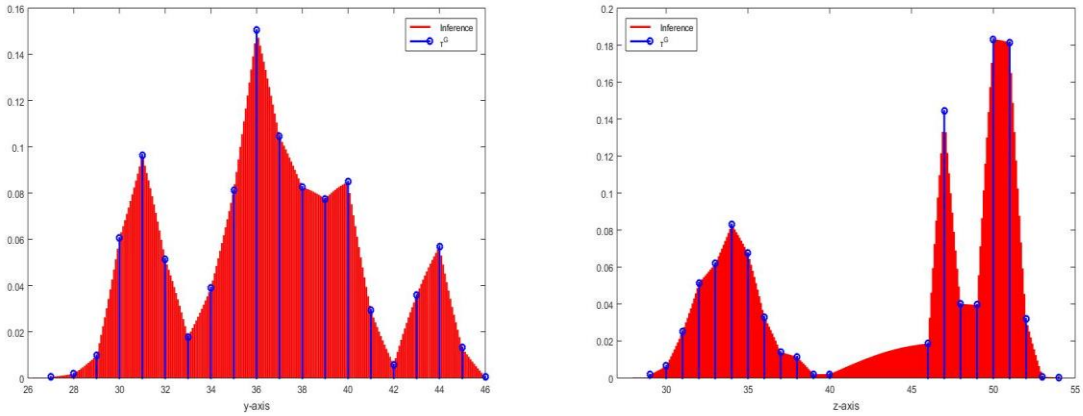


(a) Climate dataset (b) Wearable sensors dataset

Fig. 12 *3D*-examples of estimation of the *global typicality* for real climate dataset [19] and the wearable sensors dataset [23] (blue points –real data; red – estimations at hypothetical points)



(a) Climate dataset (temperature) [19] (b) Wearable sensors dataset (*x*-axis) [23]



(c) Wearable sensors dataset (*y*-axis) [23] (d) Wearable sensors dataset (z-axis) [23]

Fig.13 Examples of *2D global typicality* graphs (blue points –real data; red – estimations at hypothetical points)

Quite logical and trivial, but, importantly, we did not make any assumption about the amount of data, their (in)dependence and generation model. We only selected the distance metric (in this case, Euclidean). Moreover, in this case (because the amount of data is small and does not require recursive calculations) we did not even calculate the mean and standard deviation. The traditional approach will require a number of assumptions to be made and the *pdf* that can still be build will have non-zero values for many different values of the variable, x for which we do not know where the feasibility region is. If apply the proposed in this paper EDA approach, we will not make these paradoxical conclusions/generalisations, but will be firmly based on the observed experimental data plus the feasible hypothetical values of the variable, x at which we may decide to estimate the *global typicality* or other ensemble data properties as per EDA. It is also obvious that the results if use EDA will be identical to the well-known frequentistic probability [1-4], see Fig. 10.

Some more interesting examples of estimation of the *global typicality* based on the climate [19] and wearable sensors datasets [23] that were considered earlier are shown in Fig. 11 (*2D*) and Fig. 12 (*3D*).

It has to be stressed that the estimation is made for one hypothetical data point at a time to avoid accumulation of errors. However, if we repeat estimation many times (estimating the value of the *global typicality* for many hypothetical data points) we can get a *global typicality* graph that looks like continuous (it is not continuous because the total number of data points both existing and hypothetical is countable), see Fig.13.

IV. Properties of the EDA Operators

The typicality resembles the well-known traditional *pdf* and histograms and has itself the following additional properties:

a) it sums up to *1*;

b) its value is within the range $[0;1]$;

c) is provided in a closed analytical form, equation (27).

Fig.14 presents a visual comparison for the *2D* case between the histogram, Gaussian mixture *pdf* and the proposed *global typicality*, τ^G. The *global typicality*, contrary to histogram, in these examples, strengthens the most typical values. Moreover, by predefining the Gaussian mixture model, one can oversimplify the existing

data distribution that results in inappropriate estimation. Besides, in order to use Gaussian mixture model, one needs to select somehow the number of components (fixed number of components as in ARD-EM [24], or nonparametric methods [25]); however, this problem does not exist in the proposed approach.

One interesting property of the *density*, *D* is that for the case when Euclidean type of distance is used it takes a form of the well-known Cauchy type *pdf* , except for the fact that normalisation is needed to ensure integration to 1 [1-3]:

$$D_k\left(\boldsymbol{x}_i\right)=\frac{1}{1+\frac{\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)}{\sigma_k^2}};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{34}$$

That is, for the case of Euclidean type distance:

$$pdf_k^C(\boldsymbol{x}_i)=\frac{\Gamma\left(\frac{p+1}{2}\right)}{\pi^{\frac{p+1}{2}}\sigma_k^p}\left(D_k(\boldsymbol{x}_i)\right)^{\frac{p+1}{2}};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{35}$$

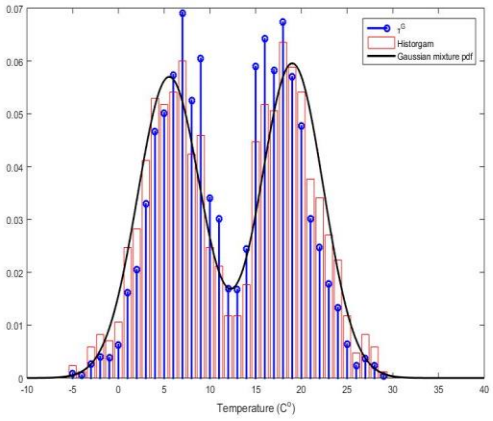
where π is the well-known constant, $\Gamma(\cdot)$ is the gamma function.

Furthermore, if Mahalanobis type of distance is used it is also of Cauchy type, and a normalisation operation is needed to turn it into the well-known Cauchy type *pdf*:

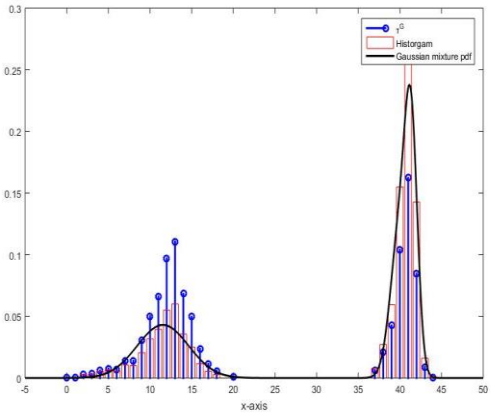
$$D_k\left(\boldsymbol{x}_i\right)=\frac{1}{1+\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}p^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{36}$$

That is,

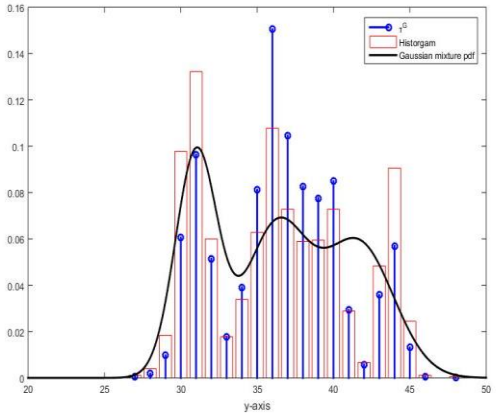
$$pdf_k^C(\boldsymbol{x}_i)=\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\pi^{\frac{p}{2}}p^{\frac{p}{2}}\left(\det\left(\boldsymbol{\Sigma}_k\right)\right)^{\frac{1}{2}}}\left(D_k\left(\boldsymbol{x}_i\right)\right)^{\frac{p+1}{2}};\quad \boldsymbol{x}_i\in\aleph_k;\quad k=\left|\aleph_k\right|>1 \tag{37}$$



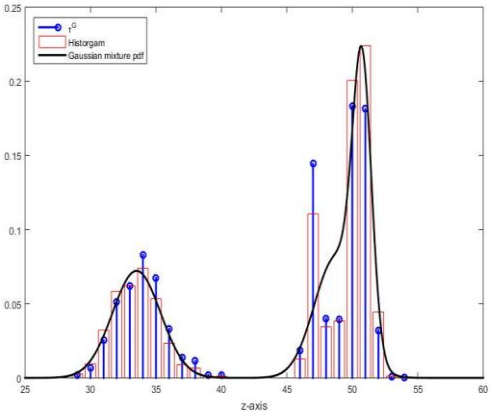
(a) Climate dataset (temperature) [19]



(b) Wearable sensors dataset (*x*-axis) [23]



(c) Wearable sensors dataset (*y*-axis) [23]



(d) Wearable sensors dataset (z-axis) [23]

Fig. 14 The comparison between histogram, Gaussian mixture *pdf* and *global typicality* for the same real climate dataset [19] and the wearable sensors dataset [23] (2D plots).

Furthermore, and interesting property of the *standardised eccentricity* if one use Euclidean type distance is that the well-known pdf can be defined through the *standardised eccentricity* as follows:

$$pdf_k^G(\boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{\sigma_k^2 - \varepsilon_k(\boldsymbol{x}_i)}{2\sigma_k^2}} \quad ; \quad \boldsymbol{x}_i \in \aleph_k ; \quad k = \left|\aleph_k\right| > 1 \tag{38}$$

Indeed, it can be shown that:

$$\sigma_k^2 - \varepsilon_k\left(\boldsymbol{x}_i\right) = -\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right); \quad \boldsymbol{x}_i \in \aleph_k ; \quad k = \left|\aleph_k\right| > 1 \tag{39}$$

V. Naïve Typicality-based EDA Classifier

Naïve Bayes classifiers are well known [1]-[3]. They perform classification based on the dominant per class likelihood expressed by a pre-defined (usually, Gaussian) *pdf*. In this paper, we borrow this concept and introduce naïve Typicality based EDA classifier (T-EDA) built on the basis of data distribution with their *global typicality* instead of a pre-defined smooth (but idealized) *pdf*.

Assuming we have C classes at time instance k and let us have the *global typicality* per class, $\tau^{G}_{k,j}$ ($j=1,2,\ldots,C$), for data sample $\boldsymbol{x}_{k+1}$, its label is given according to the following equation:

$$label\left(\boldsymbol{x}_{k+1}\right)=\arg\max_{j=1}^{C}\left(\tau^{G}_{k,j}\left(\boldsymbol{x}_{k+1}\right)\right) \tag{40}$$

Here, the *global typicality* of the v^{th} class is calculated by the following equation:

$$\tau^{G}_{k,v}\left(\boldsymbol{x}_{k+1}\right)=\frac{f_{v,l_v+1}S^{u}_{k,v}\left(\boldsymbol{x}_{k+1}\right)}{\sum_{j=1}^{C}\sum_{i=1}^{l_j}f_{j,i}S^{u}_{k,j}\left(\boldsymbol{u}_{j,i}\right)} \tag{41}$$

where the index l_v indicates the number of unique points in the v^{th} class;

$S^{u}_{k,v}\left(\boldsymbol{x}_{k+1}\right)=\left(\pi^{u}_{k,v}\left(\boldsymbol{x}_{k+1}\right)\right)^{-1}$ and f_{v,l_v+1} are calculated per class.

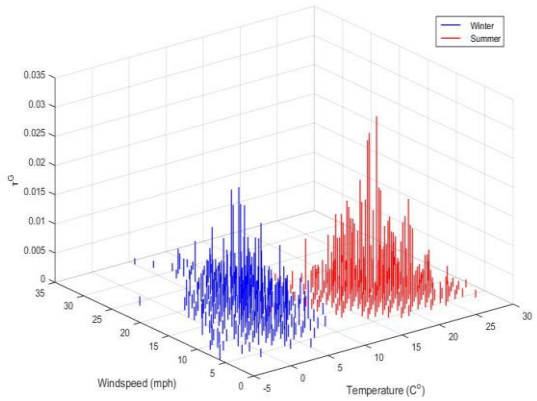
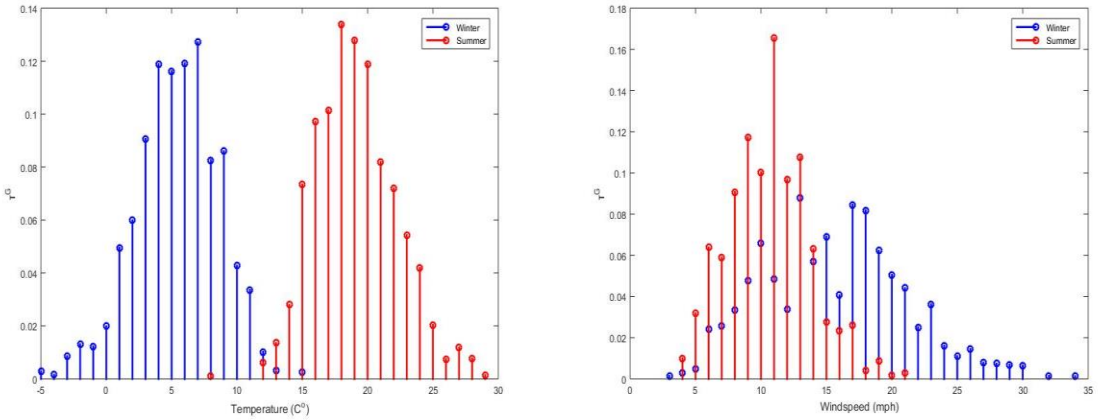


Fig. 15 *3D* plot of the *global typicality* (wind speed, temperature and the value of τ^{G}) for the same data with the two classes (winter and summer) shown with different colour.

That is, the class label for the data sample, x is the one that is most likely (has higher value of τ_v). This simple, but effective principle is the same as the one used in Naïve Bayes classifier.

The performance of the proposed naïve T-EDA classifier is further tested on a well-known challenging problem called PIMA dataset [26]. The performance of the proposed naïve T-EDA classifier was compared against the naïve Bayes, SVM[27] , eClass0[28] and Simpl_eClass0[29] classifiers. First, *90%* (*691* points) of the data set were used for training. In this paper we use the following attributes:



(a) Temperature (b) Windspeed

Fig. 16 A *2D* plot of τ vs the variable (temperature or the wind speed) for the same climate data [19]. The two classes (winter and summer) are shown with different colour. Left plot, (a): temperature, and right plot, (b): wind speed.

1) number of times pregnant;

2) plasma glucose concentration a *2* hours in an oral glucose tolerance test (mg/dl);

3) diastolic blood pressure (mm Hg);

4) triceps skin fold thickness (mm);

5) body mass index (weight in kg/(height in m)^2);

6) diabetes pedigree function.

The results are tabulated in Table 2 in the form of a confusion matrix. The proposed naïve T-EDA classifier provides *79.2%* accuracy compared with the *77.9%* of the naïve Bayes classifier, *76.6%* for the SVM classifier [27], *76.6%* for the SVM classifier [27], *58.4%* for the eClass0 classifier [28] and *63.6%* for the Simpl_eClass0 classifier [29], Fig. 17. The performance of the proposed naïve T-EDA classifier overcomes the alternative methods and it shows the capacity of solving complicated problems avoiding the need for unrealistic assumptions, restrictions, *prior* knowledge.

Table 2 Confusion Matrix for the Validation Data

Methods	**Actual\Classification**	**Negative**	**Positive**
Naïve T-EDA Classifier	Negative	76.1% (35 Samples)	23.9% (11 Samples)
	Positive	**19.4% (5 samples)**	**80.6% (26 samples)**
Naïve Bayes classifier	Negative	82.6% (38 samples)	17.4% (8 samples)
	Positive	29.0% (9 samples)	71% (22 samples)
SVM Classifier	Negative	73.9% (34 samples)	26.1% (12 samples)
	Positive	**19.4% (6 samples)**	**80.6% (25 samples)**
eClass0 Classifier	Negative	67.4% (31 samples)	32.6% (15 samples)
	Positive	54.8% (17 samples)	45.2% (14 samples)
Simpl_eClass0 classifiers	Negative	73.9% (34 samples)	26.1% (12 samples)
	Positive	54.8% (17 samples)	45.2% (14 samples)

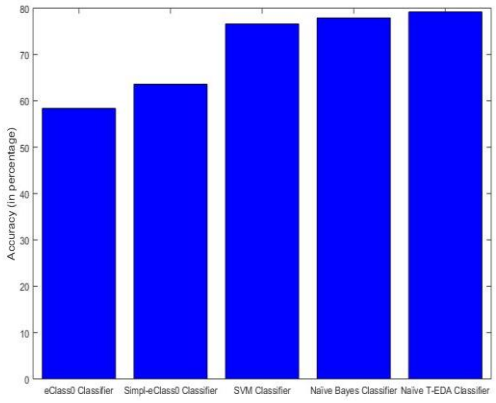


Fig. 17 Overall performances of the five classification approaches

VI. Conclusion and Future Direction

In this paper, we propose an approach to data analysis which is based entirely on the empirical observations of discrete data samples and the relative proximity of these points in the data space. At the core of the proposed new approach is the *typicality* - an empirically derived quantity which resembles probability. This non-parametric measure is a normalised form of the *square centrality*. It is also closely linked to the *cumulative proximity* and *eccentricity*. In this paper, we introduce and study two types of *typicality*, namely local and global versions. The *local typicality* resembles the well-known pdf, probability mass function and fuzzy set membership but differs from all of them. The *global typicality*, one the other hand, resembles well-known histograms but also differs from them. A distinctive feature of the proposed new approach, EDA is that it is not limited by restrictive impractical *prior* assumptions about the data generation model as the traditional probability theory and statistical learning approaches are. Moreover, it does not require an explicit and binary assumption of randomness or determinism of the empirically observed data, their independence or even their number which can be as low as couple of data samples. The *typicality* is considered as a fundamental quantity in the pattern analysis and is derived from the data directly and in a discrete form in a contrast to the traditional approach where a continuous *pdf* is assumed *a priori* and estimated from data afterwards. The *typicality* introduced in this paper is free from the paradoxes of the *pdf*. *Typicality* is objectivist while the fuzzy sets and the belief-based branch of the probability theory are subjectivist. The *local typicality* is expressed in a closed analytical form and can be calculated recursively; thus, computationally very efficiently. The other non-parametric ensemble properties of the data introduced and studied in this paper, namely, the *square centrality*, *cumulative proximity* and *eccentricity* can also be updated recursively for various types of distance metrics. Indeed, different metrics can be used, not only the usually used ones.

In short, the EDA operators as introduced in this paper, have the following properties:

- ✓ They are entirely based on the empirically observed experimental data and their mutual distribution in the data space;
- ✓ They do not require any user- or problem-specific thresholds and parameters to be pre-specified;
- ✓ They do not require any model of data generation to be assumed (random or deterministic);

- ✓ The individual data samples (observations) do not need to be independent or identically distributed; on the contrary, their mutual dependence is taken into account directly through the mutual distance between the data points/samples;
- ✓ They also does not require infinite number of observations and can work with as little as *2* data samples;
- ✓ They are free from some well-known paradoxes of the traditional probability theory;
- ✓ They can be calculated recursively for many types of distance metrics;

In addition, a new type of classifier called naïve Typicality-based EDA class is introduced which is based on the newly introduced *global typicality.* This is only one of the wide range of possible applications of EDA including, but not limited for anomaly detection, clustering, classification, control, prediction, control, rare events analysis, etc. which will be the subject of further research.

Acknowledgements

This work was supported partially byThe Royal Society grant IE141329/2014 "Novel Machine Learning Paradigms to address Big Data Streams".

References

[1] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, 2009, ISBN-13: 978-0387952840.

[2] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2007, ISBN-13: 978-0387310732.

[3] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification – 2nd Ed., Wiley-Interscience, Chichester, West Sussex, UK, 2000.

[4] T. Bayes, An Essay Towards Solving a Problem in the Doctrine of Chances, Philosophical Transactions of the Royal Society, London, England, v.53, p.370, 1763.

[5] A. N. Kolmogorov, Teoriya Veroyatnostey (Probability Theory) in Russian; published in English in 1963, American Mathematical Society, Providence, RI and republished in three volumes by the MIT Press, Cambridge, MA, reprinted 1999.

[6] V. Vapnik, R. Izmailov, Statistical Inference Problems and Their Rigorous Solutions, In A. Gammerman, V. Vovk, H. Papadopoulos (Eds.), Statistical Learning and Data Sciences, ISBN 978-3-319-17090-9, Lecture Notes in Computer Sci., v.9047, Springer, DOI:10.1007/978-3-319-17091-6_2, pp.33-71, April 2015, p.34.

[7] P. Del Moral, Non Linear Filtering: Interacting Particle Solution, Markov Processes and Related Fields, vol. 2 (4), pp. 555–580, 1996.

[8] J. Principe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Springer, 2010, ISBN: 978-1-4419-1569-6.

[9] L. A. Zadeh, Fuzzy sets, Information and Control, v. 8 (3): 338–353, 1965.

[10] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976, ISBN 0-608-02508-9.

[11] M.-Y. Chen, D. A. Linkens, "Rule-base self-generation and simplification for data driven fuzzy models." Fuzzy Systems. The 10th IEEE International Conference on. Vol. 1. IEEE, 2001.

[12] P. Angelov, R. Yager, A New Type of Simplified Fuzzy Rule-based Systems, International Journal of General Systems, v.41 (2): 163-185, Jan. 2012.

[13] L. C. Freeman, Centrality in Networks: I. Conceptual Clarification, Social Networks, v.1, 215-239, 1979.

[14] G. Sabidussi, The Centrality Index of a graph, Psychometrika, v.31, S81-603, 1966.

[15] P. Angelov, Typicality Distribution Function – A New Density-based Data Analytics Tool, International Joint Conference on Neural Networks, IJCNN, 12-16 July, 2015, Killarney, Ireland, IEEE Press, ISBN 978-1-4799-1959-8/15, DOI: 10.1109/IJCNN.2015.7280438, pp.1-8.

[16] P. Angelov, Anomaly Detection based on Eccentricity Analysis, 2014 IEEE Symposium Series in Computational Intelligence, IEEE Symposium on Evolving and Autonomous Learning Systems, EALS, SSCI 2014, Orlando, FL, USA, 9-12 Dec. 2014, pp.1-8, ISBN 978-1-4799-4495-8.

[17] P. Angelov, Autonomous Learning Systems: From Data Streams to Knowledge in Real time, John Willey, Dec.2012, ISBN: 978-1-1199-5152-0.

[18] J. G. Saw, M.C.K. Yang, and T. C. Mo, Chebyshev Inequality with Estimated Mean and Variance, The American Statistician, Vol.38 (2), 130-132, 1984, DOI: 10.1080/00031305.1984.10483182.

[19] http://www.worldweatheronline.com, accessed on 15 July 2016.

[20] D. Kangin, P. Angelov, and J. A. Iglesias. "Autonomously evolving classifier TEDAClass." Information Sciences, vol.366, p.1-11, 2016.

[21] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions." In *Artificial intelligence and Statistics*, vol. 2001, pp. 27-34. Waltham, MA: Morgan Kaufmann, 2001.

[22] P. Angelov, X. Gu, J. Principe and D. Kangin, "Empirical data analysis: A new tool for data analytics", in IEEE International Conference on Systems, Man and Cybernetics, 9-12 October 2016, Budapest, Hungary, in press.

[23] http://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer, accessed on 15 July 2016 .

[24] D. P. Vetrov, D. A. Kropotov and A. A. Osokin, "Automatic Determination of the Number of Components in the EM Algorithm of Restoration of a Mixture of Normal Distributions", Computational Mathematics and Mathematical Physics, 50(4), 733-746, 2010.

[25] J. D. McAuliffe, D. M. Blei and M. I. Jordan, "Nonparametric empirical Bayes for the Dirichlet process mixture model". Statistics and Computing, Vol. 16(1), 5-14,2006

[26] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes, accessed on 15 July 2016

[27] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods, Cambridge University Press, 2000, ISBN: 0521780195 (hb).

[28] P. Angelov and X. Zhou, "Evolving fuzzy-rule based classifiers from data streams," *IEEE Transactions on Fuzzy Systems*, vol. 16(6), p. 1462–1474, 2008.

[29] R. D. Baruah and J. Andreu, "Simpl _ eClass : Simplified Potential-free Evolving Fuzzy Rule-Based Classifiers," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC),* Anchorage, AK, pp. 2249–2254, 2011.