Typicality Distribution Function – A New Density-based Data Analytics Tool

Plamen Angelov[1,2]

[1]Data Science Group
School of Computing and Communications
Lancaster University
Lancaster, LA1 4WA, UK
p.angelov@lancaster.ac.uk

[2]Chair of Excellence
Carlos III University
Madrid
Spain

Abstract*—In this paper a new density-based, non-frequentistic data analytics tool, called typicality distribution function (TDF) is proposed. It is a further development of the recently introduced typicality- and eccentricity-based data analytics (TEDA) framework. The newly introduced TDF and its standardized form offer an effective alternative to the widely used probability distribution function (pdf), however, remaining free from the restrictive assumptions made and required by the latter. In particular, it offers an *exact* solution for *any* (except a single point) amount of non-coinciding data samples. For a comparison, that the well developed and widely used traditional probability theory and related statistical learning approaches require (theoretically) an infinitely large amount of data samples/ observations, although, in practice this requirement is often ignored. Furthermore, TDF does not require the user to pre-select or assume a particular distribution (e.g. Gaussian or other) or a mixture of such distributions or to pre-define the number of such distributions in a mixture. In addition, it does not require the individual data items to be independent. At the same time, the link with the traditional statistical approaches such as the well-known "$n\sigma$" analysis, Chebyshev inequality, etc. offers the interesting conclusion that without the restrictive prior assumptions listed above to which these traditional approaches are tied up the same type of analysis can be made using TDF automatically. TDF can provide valuable information for analysis of extreme processes, fault detection and identification were the amount of observations of extreme events or faults is usually disproportionally small. The newly proposed TDF offers a non-parametric, closed form *analytical* (quadratic) description extracted from the real data realizations *exactly* in contrast to the usual practice where such distributions are being pre-assumed or approximated. For example, so called particle filters are also a non-parametric approximation of the traditional statistics; however, they suffer from computational complexity and introduce a large number of dummy data. In addition to that, for several types of proximity/similarity measures (such as Euclidean, Mahalonobis, cosine) it can be calculated recursively, thus, computationally very efficiently and is suitable for real time and online algorithms. Moreover, with a very simple example, it has been illustrated that while traditional probability theory and related statistical approaches can lead in some cases to paradoxically incorrect results and/or to the need for hard prior assumptions to be made. In contrast, the newly proposed TDF can offer a logically meaningful result and an intuitive interpretation automatically and exactly without any prior assumptions. Finally, few simple univariate examples are provided and the process of inference is discussed and the future steps of the development of TDF and TEDA are outlined. Since it is a new fundamental theoretical innovation the areas of applications of TDF and TEDA can span from anomaly detection, clustering, classification, prediction, control, regression to (Kalman-like) filters. Practical applications can be even wider and, therefore, it is difficult to list all of them.

Keywords—TEDA, typicality, eccentricity, data density, pdf, non-parametric data distributions.

I. INTRODUCTION

Traditional probability theory [1], including the widely celebrated Bayesian approach [2], which were introduced two-three centuries ago are based on the frequentistic approach to represent uncertainties and make a number of strong assumptions, which usually do not hold in practice. These include the requirements to have a theoretically infinite (or practically, for an approximation, a very large) amount of observations (data samples), these data samples to be independent, etc. They are well developed tools to address the "pure" random variables and processes for which they were designed in the first place, such as gambling, games, etc. The basic frequentistic concept was later developed extensively into a variety of methods and approaches. In order to apply them to real processes of interest (such as climate, economical, social, mechanical, electronic, biological, etc.), however, the vast majority of them rely on *prior* assumption of smooth and monotonic distributions, such as Gaussian/normal, Cauchy, etc.[2],[3] or a mixture of them [4]. If use a mixture of (Gaussian) distributions the question arises: how to determine the modes of the pdf and, respectively, the number of functions in the mixture. This is usually done offline by the human user and as a result of approximations (not exact) which poses further questions and problems. The more recent alternative is to approximate the distributions using non-parametric, data-centered functions, such as particle filtering [5] or the entropy-based information-theoretic learning [6] methods. However, they do not depart completely form the Gaussian assumptions which are used for describing the distribution around the data points.

Nowadays, the demand is growing for new concepts in Data Analytics that are centered at the data rather than at theoretical *prior* assumptions which are then being confronted with the real experimental data. The latter was a dominating

trend in the last couple of centuries, but it is being increasingly shifted towards a data-centric approach lately. Nowadays, with the ubiquituous spread of data in nearly every form of human activity it is of significant interest to have tools and framework/concept to extract the inherent data pattern rather than to simply try to fit it to the template of an assumed *a priori* distribution.

The first step in establishing a systematic theoretical framework that is entirely data-driven and makes no prior assumptions was the introduction of TEDA (the typicality and eccentricity based data analytics framework) in 2014 [7],[8]. In the present paper TEDA is further developed by introducing the TDF (typicality distribution function) as an effective alternative to the well-known probability density function (pdf) [2],[3] and membership functions of fuzzy sets [9]. TDF is entirely data-driven and does not require any *prior* restrictive assumptions to be made unlike the traditional pdf, membership functions and non-parametric approaches such as particle filters, information-theoretic learning etc. Moreover, it can be calculated recursively and computationally very efficiently for Euclidean, cosine, Mahalonobis and Manhattan type distance metrics. Simple examples are provided mainly to illustrate the concept while the further developments of the theory to design new type of anomaly detection, clustering, classification, prediction, regression, control, filtering approaches and applications to various fields is left for future publications due to the space and time limitations.

The remainder of the paper is organized as follows: section II provides a brief introduction of the basic concepts of TEDA; section III introduces new TDF, and its standardized version, $\overline{m}$ and provides the mechanism for inference; section IV provides some simple examples of TDF and compares them with the pdfs; and finally, section V concludes the paper with directions of the future work and applications.

II. INTRODUCTION TO TEDA

TEDA was introduced in 2014 [7],[8] aiming to offer a fully data-driven and *prior*-assumptions-free framework for Data Analytics. It is based on the data density rather than on frequency of occurrence assumed distributions [2],[3] or on subjective judgment [9] as its predecessors were. It, therefore, does not require *any prior* assumptions to be made, such as for example:

a) independence of the individual data items from each other;
b) large (theoretically, infinite) number of data items;
c) *prior* assumption of the distribution or kernel (most often, normal/Gaussian).

Indeed, *real* processes (e.g. climate, economic, physical, biological, social, psychological, etc.) which are of practical interest are often complex and uncertain, but they are **not** *purely* random; they **do have** inter-sample dependence, not necessarily normal/Gaussian distributions and definitely not infinite number of observations. It is a well-known fact that statistical approaches (and probability theory) does not work (well) on small amount of data. However, for many important problems such as extreme events analysis and predictions (e.g. climate, earthquakes, etc.), fault detection the amount of data (for the faulty cases) can be very small.

TEDA which was introduced recently [7]-[8] offers an efficient alternative to the traditional statistical and probabilistic framework (as well as to the fuzzy set theory). At the same time, it can also be seen as an augmentation of both and can work with any ***real*** data with as little as a couple of data samples. The only exception is if all data samples coincide in a single point; for such a singular case both TEDA and TDF, in particular, are not defined. In addition to such singular case, for **pure** random variables and processes (such as gambling, games, e. g. throwing dices, tossing coins, selecting balls from bowls, etc.) the traditional probability theory is best fitted, indeed. We can summarize the areas where the traditional probability theory (and statistics) is best fitted and the area covered better by TEDA are as follows:

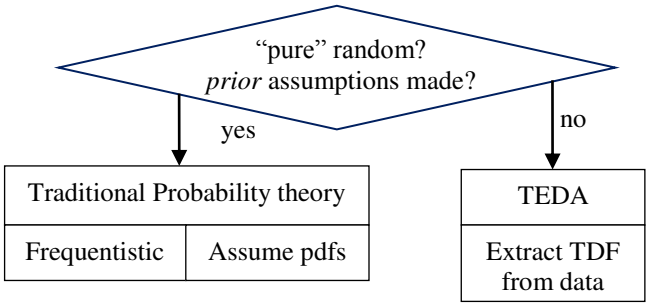

Fig. 1 Areas for which traditional probability theory and TEDA are best fitted

Let us consider the data space $\mathfrak{X} \in \mathbb{R}^n$, which consisit of *n*-dimensional data points. Within this space, we can define the distance $\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{y})$, which can be, for example, Euclidean, Mahalanobis, cosine, $\boldsymbol{L_1}$, or any other. Then, let us consider the data points as an ordered sequence $\{x_1, x_2, ..., x_k, ...\}$ $x_i \in R^n$ $i \in N$ where the index $\boldsymbol{k}$ may have the physical meaning of time instant when the data item has arrived. For this reason, $\boldsymbol{k}$ will be referred as time instant further for simplicity. Within TEDA we consider [7],[8]:

- *accumulated proximity,* π from a particular, j^{th}, $j>1$ data point $\boldsymbol{x} \in \mathfrak{X}$, to all remaining, $k>1$ data points:

$$\pi_k(x_j) = \pi_{jk} = \sum_{i=1}^{k} d_{ij} \quad k > 1 \tag{1}$$

where d_{ij} denotes a distance between data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

- *eccentricity* of the j^{th} data item calculated when $k>2$ non-identical data items are available:

$$\xi_{jk} = \frac{2\pi_{jk}^{k}}{\sum_{i=1}^{k} \pi_{ik}} \quad \sum_{i=1}^{k} \pi_{ik} > 0 \quad k > 2 \tag{2}$$

where π_{jk}^{k} denotes *accumulated proximity,* π from a particular, j^{th}, $j>1$ data point $\boldsymbol{x} \in \mathfrak{X}$, to all remaining, $k>1$ data points when $k>2$ non-identical data items are available.

These quantities (π and ξ) can be defined either ***locally*** (for a part of) or ***globally*** (for all data points) and can be calculated recursively for certain type of distances. For example, if we use cosine distance normalized to be within the range *[0;1]*, we come to the following expressions [8]:

$$\pi_j = \sum_{i=1}^{k} \frac{1}{2}\left(1 + \frac{x_i x_j}{\|x_i\|\|x_j\|}\right) = \frac{k}{2}\left(1 + \bar{\mu}_k \frac{x_j}{\|x_j\|}\right); \quad \|x_j\| \neq 0 \tag{3}$$

where

$$\bar{\mu}_k = \frac{1}{k}\sum_{i=1}^{k} \frac{x_i}{\|x_i\|} \quad \|x_i\| \neq 0 \tag{4}$$

or recursively

$$\bar{\mu}_k = \frac{k-1}{k}\bar{\mu}_{k-1} + \frac{1}{k}\frac{x_k}{\|x_k\|} \quad \|x_k\| \neq 0 \tag{5}$$

$$\sum_{i=1}^{k} \pi_{ik} = \sum_{i=1}^{k} \frac{1}{2}\left(1 + k\bar{\mu}_k \frac{x_j}{\|x_i\|}\right) = \frac{k}{2}\left(1 + k\bar{\mu}_k \bar{\mu}_k\right); \quad \|x_j\| \neq 0 \tag{6}$$

The eccentricity can be determined by:

$$\xi_{jk} = \frac{1 + \bar{\mu}_k \frac{x_j}{\|x_j\|}}{1 + k\bar{\mu}_k \bar{\mu}_k}; \quad \|x\| \neq 0 \tag{7}$$

If use Euclidean distance one gets [8]:

$$\pi_{jk} = k\left(\| x_j - \mu_k \|^2 + X_k - \| \mu_k \|^2\right) \tag{8}$$

$$\mu_k = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k \quad \mu_1 = x_1 \tag{9}$$

$$X_k = \frac{k-1}{k}X_{k-1} + \frac{1}{k}\| x_k \|^2 \quad X_1 = \| x_1 \|^2 \tag{10}$$

$$\sum_{i=1}^{k} \pi_{ik} = \sum_{i=1}^{k-1} \pi_{i(k-1)} + 2\pi_{kk} \quad \pi_{11} = 0 \tag{11}$$

where μ - recursively updated (local or global) mean;
X is the recursively updated squared norm sum.

Further, in TEDA a condition which provides ***exactly*** the same result as the so-called Chebyshev inequality [12] without making any assumptions on the amount of data and their independence was introduced for Euclidean distance [8],[11]:

$$\xi_k > \frac{n^2+1}{k} \tag{12}$$

which can be called the TEDA ***eccentricity inequality***.

Similar (not the same, but for the case of Mahalonobis type distance subject to a coefficient represneting the dimensionality of the data vector, ***x,*** [11]) inequalities can also be derived for other types of distances, such as Mahalonobis [11], cosine, L_1, In the above expression, n is the well-known factor from the so-called ***"$n\sigma$"*** principle (where σ denotes the standard deviation). As a reminder [2],[3], this principle guarantees that for normally (Gaussian) distributed random variable and a representatively large amount of data the vast majority of the data (>99.7% if use n=3) can be considered "normal" and the probability for a data item to be abnormal (further than 3σ away from the mean, μ) is, respectively <0.3%. A more general property is given by the so-called Chebyshev inequality [12] mentioned above. Namely, for *any* distribution having a ***large amount*** of ***independent*** data points the probability for a data point to be $>n\sigma$ away from the mean, μ is $<1/n^2$. For example, the probability to have a data point distant form the mean more than, 3σ is *<1/9* (or *~11%*). Aiming to avoid creating too many false positives they also use in practice 6σ or even higher n to guarantee that *<1/36* (or *~3%*) of the data are declared anomalous [12].

III. TDF

A. Standardized eccentricity, ς

In this paper, the TEDA is further developed by introducing TDF. Let us start by analyzing the expression for the eccentricity, (2) and the *TEDA eccentricity inequality*, (12). In this paper, we introduce standardized eccentricity as follows:

$$\varsigma_{jk} = \frac{2\pi_{jk}^{k}}{\bar{\pi}_k} \quad \bar{\pi}_k > 0 \quad k > 2 \tag{13}$$

where $\bar{\pi}_k = \frac{1}{k}\sum_{i=1}^{k}\pi_{ik}$ is the average accumulated proximity, π from a given point to all other points.

Please, note the difference between the Greek symbols ξ, ζ, and ς which represent, respectively the eccentricity (equation (2)), normalized eccentricity [8], $\zeta = \xi/2$ and ς. The latter can also be expressed as follows:

$$\varsigma_{jk} = k\xi_{jk} \tag{14}$$

It can easily be seen that ς has some very interesting properties. For example, it is very suitable so called Big Data problems when k can be *very large* and both ξ and ζ can potentially lead to computational problems (hardware dependent, not theoretically restrictive). ς (see (13) and (14)) is free form such problems. For normlaised data the distances are limited to *1* and $\bar{\pi}_k$ can be updated through an expression similar to (9):

$$\bar{\pi}_k = \frac{k-1}{k}\bar{\pi}_{k-1} + \frac{2}{k}\pi_{kk} \quad \pi_{11} = 0 \tag{15}$$

or by learning [13]-[15] using a learning rate, α ($0<\alpha<1$):

$$\bar{\pi}_k = (1-\alpha)\bar{\pi}_{k-1} + \alpha\pi_{kk} \quad \pi_{11} = 0 \tag{16}$$

Obviously, if $\alpha=1/k$ equation (16) reduces to (15) but in order to avoid the problems with large k one can select any value of α between *0* and *1* and get an assymptotic approximation of (15). This learning process is a special case of the well known least mean squares principle and has been used widely in machine learnrng literature [14]-[16].

The meaning of ς is that of a comparator between the accumulated proximity, π from a given point with the average accumulated proximity, $\overline{\pi}$ of all data points. The values of ς are positive but can be >1 when a point is more than one σ away from the mean, μ. That is, we can redefine the *TEDA eccentricity inequality* and discover anomalies by analyzing ς:

$$\varsigma_{jk} > n^2+1 \tag{17}$$

or equivalently:

$$2\pi_{jk}^{k} > (n^2+1)\overline{\pi}_k \tag{18}$$

Equation (17) can be called *TEDA standardized eccentricity inequality* and (18) can be called *TEDA accumulated proximity inequality*. Not only they look simpler and are more convenient to use (the latter one even does not have a division) but for large k they are much more suitable.

If we analyze further the standardized eccentricity, ς we can see that for the vast majority of the data (as described above) the values of ς lie in the range $]0;\ n^2+1]$ and only for less than $1/n^2$ of the data it will have a value bigger than (n^2+1). Moreover, this conclusion ***does not require any prior assumption*** to be made about the type of the distribution of the data or independence of the data samples or, moreover, about the number of data items/points. Indeed, it works perfectly well for as little as a couple of data points. Furthermore, in this paper we suggest an automatic way of determining the value of n as a function of the number of data points available as follows:

$$n=\begin{cases}\sqrt{k}; & k<\left(n^*\right)^2\\ n^*; & k\geq\left(n^*\right)^2\end{cases} \tag{19}$$

where n^* denotes the traditionally used values such as *3* and *6*.

B. TDF definition

Starting from the standardized eccentricity, ς that was introduce above the typicality distribution function, TDF can now be defined as follows:

$$m_{jk}=1-\frac{\varsigma_{jk}}{n^2+1} \qquad j=1,2,...,k-1,k \tag{20}$$

or equivalently

$$m_{jk}=1-\frac{2\pi_{jk}^{k}}{\left(n^2+1\right)\overline{\pi}_k} \tag{20a}$$

Obviously, for Euclidean, Mahalnobis, cosine, L_1 types of distance measures the typicality values can be calculated and updated recursively and there is no need to memorise all data points. For example, for Euclidean type distance it becomes:

$$m_{jk}=1-\frac{2}{n^2+1}\frac{\delta_{jk}^2+X_k-\|\mu_k\|^2}{\sigma_k^2+X_k-\|\mu_k\|^2} \tag{21}$$

where $\delta_{jk}^2=\|x_j-\mu_k\|^2$ denotes the deviation form the mean of a particular point, x_j; $\sigma_k^2=\frac{1}{k}\sum_{j=1}^{k}\delta_{jk}^2$ denotes the well-known squared standard deviation.

Let us analyze the anlaytic expression of TDF for the Eculidean distance. It is, obviously, a quadratic function of the partciular, j^{th} data point, x_j. The maximum value which this function can get is *1* when $x_j=\mu_k$. For all other values of x_j it is less than *1*. It gets exatcly *0* when standardized eccentricity is $=n^2+1$ (borderline case for a point to be considered an outlier). Obviously, it is dependent on the choice of n, but with the suggested automatic mechanism (19) it is automatic. For the minority of the cases the value of m can get negative. The probability this to take place according to the *TEDA standardized eccentricity inequality,* (17) is $<1/n^2$.

Analysing TDF we can see that the sum of m_j for all values of x_j; $j=1,2,...,k$ is always bigger than *1*:

$$\sum_{j=1}^{k}m_{jk}=k\left(1-\frac{2}{n^2+1}\right)=k\left(\frac{n^2-1}{n^2+1}\right) \tag{22}$$

On the surface, the TDF, m resembles very much fuzzy sets membership functions (having a maximum value of *1* and a sum of vlaues larger than *1*, being a smooth monotonic function, etc.) but it is quite different in nature. It can (though very rarely) have negative values.

Starting from equation (22) we standardize TDF (equation (20) or (20a)) by dividing to the factor $k\left(\frac{n^2-1}{n^2+1}\right)$ which results in a new quantity called standardized typicality distribution, $\overline{m}$:

$$\overline{m}_{jk}=\frac{m_{jk}}{k}\left(\frac{n^2+1}{n^2-1}\right)=\frac{1}{k\left(n^2-1\right)}\left(n^2+1-\varsigma_{jk}\right) \tag{23}$$

or briefly:

$$\overline{m}_{jk}=\frac{n^2+1-\varsigma_{jk}}{k\left(n^2-1\right)} \tag{23a}$$

It is now easy to check that $\overline{m}$ sums up to *1*:

$$\sum_{j=1}^{k}\overline{m}_{jk}=1 \tag{24}$$

by definition, because:

$$\overline{m}_{jk} = \frac{m_{jk}}{\sum_{i=1}^{k} m_{ik}} \tag{25}$$

Additionally, for majority of the data (probability for this is $>1-1/n^2$) it lies within the range *[0;1]*:

$$P\left(0 \leq \overline{m}_{jk} \leq 1\right) \geq 1 - \frac{1}{n^2} \tag{26}$$

with negative values being associated with outliers (which are $<1/n^2$ for any type of distribution):

$$P\left(\overline{m}_{jk} \leq 0\right) \leq \frac{1}{n^2} \tag{26a}$$

The standardized TDF, $\overline{m}$ resembles very strongly the classical pdf without, however, requiring the strong restrictive assumptions associated with the latter to be made and being negative for outliers (thus detecting them automatically). One can also argue that it is a function of n (of the choice made by selecting n), but this is not a problem- or user-specific parameter and the rationale for its choice is quite obvious and stable. Apart form this, the only other restriction/requirement is to have at least one data point that differes from all others and at least a couple of data points in a ***real***, not "*purely*" random process. No any other assumptions are necessary and no other restrictions apply.

While the values of TDF do not diretcly depend on k and for any value of k will not suffer form compuational problems, the values of the standardised TDF, $\overline{m}$ can become very small nominaly for large k. Unlike, traditional pdf which (at leats theoretically) has been defined as a continuous function with infinite number of values in the independent argument, TDF (and the whole TEDA) is data-centric and is discrete by nature (therefore, we used sum and not integral). Therfore, for plotting the values it is correct to use stem plot rather than continuous envelope curve. For large values of k it is recommendable to plot a histogram-like stem plot where the values of the independent argument are summed up in bins. Obviously, the value of $\overline{m}$ for a certain bin will be a sum of vlaues of $\overline{m}$ for all data points that fall into that bin. This is quite different form histograms used in the traditional probability theory to represent the pdf because in TDF case these are just presnetational mechanisms and not a reqirement. This way of presenting stadnardised TDF values, $\overline{m}$ is entirely optional and aimes primarily internpretation and computation convenience. The exmaples in the next section demosntrate that, in general, this is not necessary. In addition, TDF can perfectly well work with as little as couple of data samples while traditional histograms of pdf do require a large amount of data.

In the next section a number of simple illustrative exmaples will be provided and the standardized TDF, $\overline{m}$ will be compared with the traditional pdf. One of the examples will demosntrate the confusion when applying the traditional probability and pdf while the newly proposed TDF copes very well and provides a very logical result.

C. Inference using TDF

Finally, the problem of producing an inference using TDF and its standardised form, $\overline{m}$ will be considered.

Let us have a TDF and/or $\overline{m}$ derived form the data and let us try to infer the standardised typicality of a value of x that never took place. To do this, we can simply assume that the next data point, x_{k+1} is the point of interest and update the values of μ_{k+1} by equation (9), X_{k+1} by (10), $\sum_{i=1}^{k+1} \pi_{i(k+1)}$ by (11), $\overline{\pi}_{k+1}$ from (15) in relation with (8), ς_{k+1} from (13), m_{k+1} from (20) and finally, the $\overline{m}_{k+1}$ from (23a). All these derivations are non-iterative, can be done online, recursively and, thus computationally very efficiently. They will provide, as a result in vast majority of the cases a value between *0* and *1* which can be interpreted as a percentage of the standardised typicality (a likelihood). For the minority of the cases when these values will be negative the conclusion is that they are eccentric and not typical, but possible nevertheless (the probability that such values can occur is guaranteed to be $<1/n^2$ according to the TEDA standardised eccentricity inequality, (17).

IV. Illustrative Examples

Because of the space limitations in this section illustrative numerical examples will only be provided aiming primarily a proof of concept. First, several simple *1D* examples will form a TEDA/TDF primer to get started. Then a simple *1D* climate example will demonstrate that the traditional probability theory does not provide a satisfactory representation unlike TEDA and TDF or requires many hard assumptions to be made and even in such a case does provide an approximate one. Finally, couple of still simple, *1D* but real climate data-based examples will demonstrate the TDF and $\overline{m}$.

A. TDF Primer

Let us start with the basics. Let us consider the simplest possible case of just two non-coinciding data points. It is a trivial example, indeed, but for completeness, we can start with it. If we have two non-coinciding points, A and $B \neq A$ and we denote the distance between them as d then we will also obviously have $k=2$; $\pi_A = \pi_B = d$ that is $\pi_2 = \{d; d\}$; $\Sigma\pi_2 = 2d$; $\overline{\pi} = d$; $\varsigma_{j2} = 2$; that is $\varsigma_2 = \{2;2\}$; $n = \sqrt{2}$; $m_2 = \left\{\frac{1}{3}; \frac{1}{3}\right\}$; $\overline{m}_2 = \left\{\frac{1}{2}; \frac{1}{2}\right\}$. This is quite natural and expected; each of the two points (regardless of the specific position of points A and B, the type of the distance and dimensionality) is equally typical and likely (50% each in terms of $\overline{m}$), see Fig.1 (in Fig.1 we depicted a *1D* case, but the same conclusion can be made for any dimensionality). The TDF does not reach its

theoretical maximum of *1* because it can be acquired at the mean value which is between the two points.

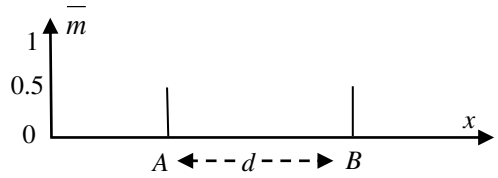

Fig. 1 A trivial example: $\bar{m}$ for 2 equally spaced data points.

As a second trivial example we can consider three points, so $k=3$; $n=\sqrt{3}$. Even with this simplistic example there are various options. For example, the three points, *A, B* and *C* may be equally distant from each other in the data space, forming a unilateral triangle:

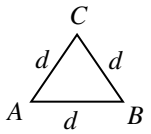

Fig. 2 An illustration of *3* equally spaced points in a *2D* space.

In such a case, $\pi_3=\{d;d;d\}$; $\Sigma\pi_3=3d$; $\bar{\pi}=d$; $\varsigma_3=\{2;2;2\}$; $m_3=\{0.2;0.2;0.2\}$; $\bar{m}_3=\left\{\frac{1}{3};\frac{1}{3};\frac{1}{3}\right\}$. This is also quite expected and natural. However, if the data are not equally spaced between themselves, for example, if we have the three points placed as depicted in Fig.3

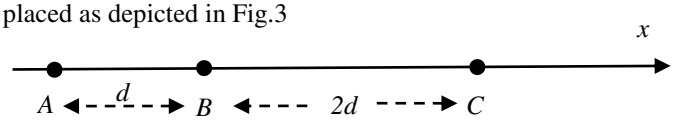

Fig. 3 A trivial *1D* example with three data points which are not equally spaced.

For this case we have $k=3$; $n=\sqrt{3}$; $\pi_3=\{4d;3d;5d\}$; $\Sigma\pi_3=12d$; $\bar{\pi}=4d$; $\varsigma_3=\{2;1.5;2.5\}$ $m_3=\left\{0.5;\frac{5}{8};\frac{3}{8}\right\}$; $\bar{m}_3=\left\{\frac{8}{24};\frac{10}{24};\frac{6}{24}\right\}$. Even from this trivial *1D* example with just *3* data points it is obvious that in TEDA (unlike in traditional probability theory) what matters is not just how often we have an observation with a certain value but also how these values are mutually distributed in the data space. For example, the point *B* is somewhat more typical while point *C* is the least typical one.

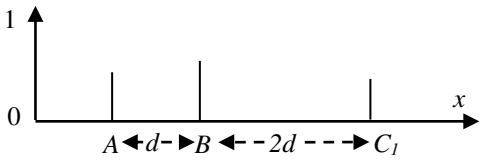

Fig. 4 $\bar{m}$ for the second trivial example

It is obvious that standardized TDF provides a different type of information about the importance and likelihood of the data points in comparison with the traditional probability theory and with the fuzzy sets. We argue that for real processes (not dices, coins and other gambling, games or *pure* random processes) this is more realistic that point *B* is more likely and more typical than point *A* or point *C*.

B. Simple 1D climate primer

This difference is even more obvious if we consider such a simple hypothetical example. Let us have five data points representing the temperature in a city. For example, we may have two cases of 10^oC and one case of 16.9^oC, 18.1^oC and 19.3^oC, respectively. The well known traditional probability theory will either suggest that the probability of having 10^oC is twice as big at *40%* (in comparison with the *20%* for each one of the other observations), Fig.5. Even if assume a distribution (e.g. of a Gaussian or other type) it needs to be parameterized (finding the mean and standard deviation). If assume a mixture of distributions it needs to be pre-determined the number of such distributions (in this simple case, may be 2) and each one of them also need to be parameterized. Instead, TEDA offers an automatic and exact (not approximate) way of calculating m and $\bar{m}$ without the need to assume/select the type of the distribution, to parameterize it or to decide if a mixture of distributions is best and how many components such a mixture should have. For such a simple example for $x=\{10;16.9;10;18.1;19.3\}^oC$ the values of $\bar{m}_5=\{0.186;0.219;0.186;0.213;0.196\}$ are depicted on Figure 5. It is clear that they are quite logical.

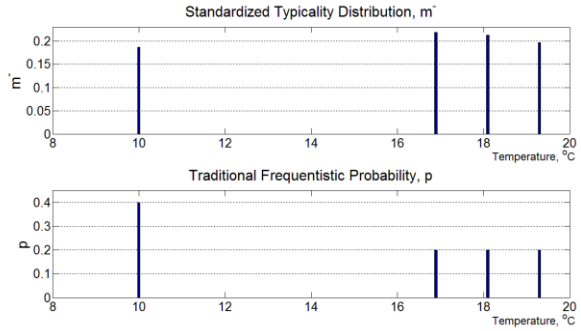

Fig. 5 $\bar{m}$ for the simple *1D* climate example.

C. Modes detection by TDF

TDF can be seen and used as an automatic mechanism for outliers/anomalies detection which does not require any *prior* assumptions to be made about the distribution, amount of data and their independence. It can be used for a screening to find outliers and, in this way, to automatically find modes of distributions and, in effect, perform clusteirng and extracting multi-modal distributions (if the data pattern requires this) without pre-defining how many modes there will be.

This process can simply start with calcuating the ***global*** *m* for each data point (offline, online or in an evolving manner). As a next step, the number fo points witin the σ-vicinity around the mean, μ can be compared with the number of points outside this vicinity. The rationale being that the majority of the points have to lie close ot the mean or, alterntaively a multi-modal reprensetation is better. If this is

not the case (if the number of points outside the σ-vicinity is larger than the number of points within the σ-vicinity of the mean, μ), e.g. as depicted in Fig.5 (where $\mu \approx 14.88$; $\sigma \approx 4$) then additional modes have to be formed. For all points which lie outside σ-vicinity of the mean, μ if the distance between them is less than σ are considered toghether and form a local mode for which a local mean, μ^i ($i=1,2,...$) is calcuated. These new local means replace the global mean, μ. For the simplistic example, depicted in Fig.5 this will result in a function with two modes around *10°C* and around *18°C* automatically derived from the data. It has to be noted that this is very logical and exatcly what a human user would probably decide to do manually. The number of points which satisfy this condition, k^i ($i=1,2,...$) is also calcuated/updated as well as the respective local quantities π^i, $\overline{\pi}^i$, ς^i, m^i, and $\overline{m}^i$ ($i=1,2,...$). In an online and evolving mode for each newcoming point the distance to all previously discovered modes of the distribution (local means) can easily be calculated. The newcoming point can then be associated with the nearest one and with the other points associated with it. In this way we can get multiple modes and data sub-sets (clusters) associated with them.

D. Simple real 1D illustrative examples

Finally, let us consider more realistic, but still quite simple, *1D* examples of climate data. In Fig.6 we depict $\overline{m}$ for the temperature during December 2014 in Central England [17].

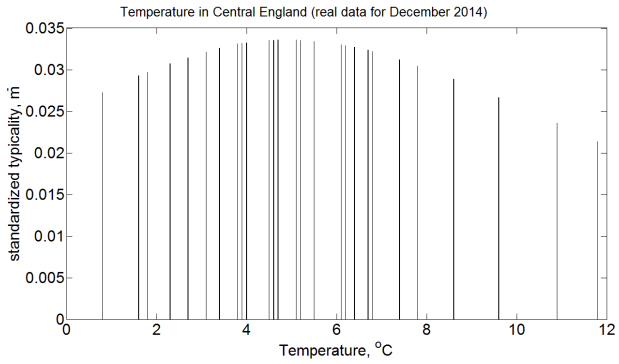

Fig. 6 $\overline{m}$ for December 2014 temperature in Central England.

Another real, yet simple illustrative example depicts in Figure 7 the January temperature in Central England for a period starting *1772* till present day.

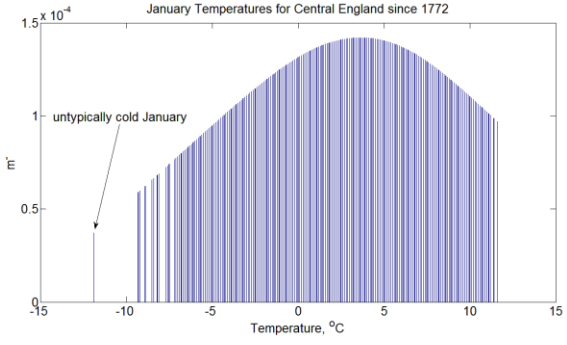

Fig. 7 $\overline{m}$ for another simple *1D*, but real, climate data example (January temperature in Central England since 1772).

It is clear that one of the days is untypically called (*-12oC*), but is not abnormal ($\overline{m}$ is positive). Let us consider A hypothetical example of a very warm December (say, *26°C*) will be extremely unlikely (the value of $\overline{m}$ will become negative indicating this untypical/eccentric case), Fig.8. In this example all but one data point are real (same as in Fig.6).

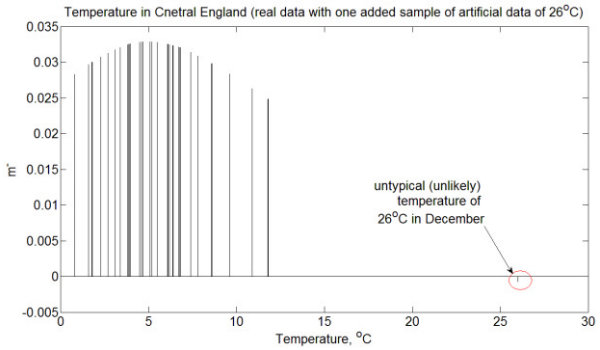

Fig. 8 A hypothetical data point mixed with the real data (an extremely warm December day with temperature 26ºC). The value of $\overline{m}$ is negative indicating this is untypical/unlikely.

Finally, real multivariate data about minimum and maximum daily temperature in Marseille, France for the period 1956-1999 are presented in Figs.9 and 10. One can see the non-Gaussian nature.

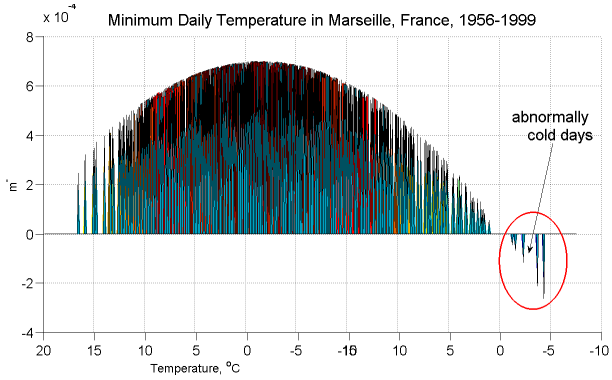

Fig. 9 Min and max daily temperature in Marseille (1956-'99)

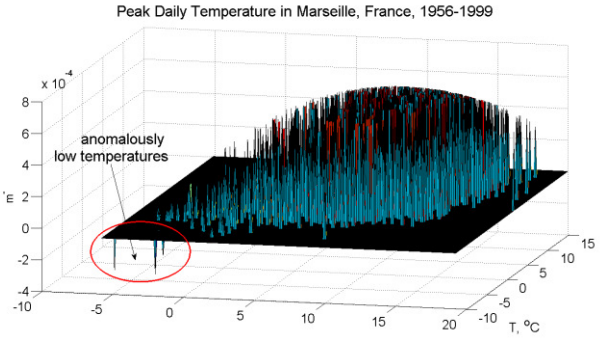

Fig. 10 A *3D* plot of $\overline{m}$ vs peak temperature in Marseille (1956-1999, real data).

V. CONCLUSIONS

In this paper, the recently introduced data analytics framework TEDA is further developed by the introduction of TDF and its

standardized form, $\overline{m}$. It offers a closed analytics (quadratic) form formulae which provides the likelihood somewhat similar but not the same to the pdf. TDF, on the other hand, offers a typicality distribution which resembles a data-derived membership function of a fuzzy set. These are based on a series of normalizations/standardizations. The proposed TDF and $\overline{m}$ are free from the restrictive assumptions made and required by the traditional probability theory and statistics. In particular, it offers exact values for any (as little as a couple) number of data points, does not require their independence (on the contrary assumes that the process is real and not *purely* random). It does not require the user to pre-select or assume smooth distributions (e.g. Gaussian or other) or a mixture of such distributions and to pre-define the number of such functions in the mixture. The importance of the good choice of prior distributions in traditional probability theory is well known. For example, in [1] it says on p.23 "…and indeed Bayesian models based on poor choice of prior can give poor results with high confidence". Without making any prior assumptions and requiring any subjective input TEDA offers a direct mechanism for calculating and updating the *typicality* as a form of representing the *likelihood* of any ***real*** variable but "*pure*" random (such as gambling, games, etc. that satisfy the strong assumptions listed in section II as a)-c) for which the traditional probability theory was actually designed and is best suited and without subjective forms of uncertainty (preferences) for which fuzzy set theory was designed and is best suited for. For all other variables (not the "*pure*" random and not subjective), the inference in TEDA provides and updates the *typicality* distribution automatically.

The newly proposed TDF can provide valuable information for analysis of extreme processes, fault detection and identification were the amount of observations of extreme events or faults is disproportionally small. At the same time, the link with the traditional statistical approaches such as the well-known "$n\sigma$" analysis, Chebyshev inequality etc. offers the interesting conclusion that without the restrictive prior assumptions listed above to which these traditional approaches are tied up the same type of analysis can be made using TDF automatically.

Since it is a new fundamental theoretical innovation the areas of applications of TDF and TEDA can span from anomaly detection, clustering, classification, prediction, control, regression to (Kalman-like) filters. Practical applications can be even wider and therefore it is difficult to list all of them.

TEDA is entirely based on the density and proximity in the data space. It is not tied up to any particular type of distance and can be recursively expressed by using a number of types of distances, such as Euclidean, Mahalonobis, cosine, Manhattan.

It was demonstrated on some very simple and intuitive real data of the temperature distribution in Central England that TDF can be generated automatically form the data without any prior assumptions and provides logical information about the typicality and likelihood of a particular value of the data through a straightforward inference. The automatic extraction of multi-modal distributions, new clustering methods and other applications (including filters, classifiers, predictors, controllers etc.) will be a matter of forthcoming publications. The problem of anomaly (also the related fault- and novelty-) detection using typicality was already described in [8] and its extension using TDF was described in this paper (see Fig. 8 for example).

VI. ACKNOWLEDGEMENTS

The author would like to acknowledge the Chair of Excellence 2015 award by Carlos III University, Madrid, Spain and Santander Bank. The author would also like to acknowledge the help of Mr. Dmitry Kangin in making the demo for the climate data and the useful discussion of the draft text.

REFERENCES

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

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

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, NY, USA, 2000.

[4] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and EM algorithm," *SIAM Review*, vol. 26, No.2, pp. 195-239, April 1984.

[5] B. Ristic, S. Arulampalam, N. Gordon *Beyond the Kalman Filter: Particle Filters for Tracking Applications,* Artech House Radar Library, 2004.

[6] J. C. Principe, *Information Theoretic Learning: Rényi's entropy and kernel perspectives*, Springer Verlag, Germany, April 2010

[7] P. Angelov, Outside the Box: An Alternative Data Analytics Framework, *Journal of Automation, Mobile Robotics and Intelligent Systems,* vol. 8, No2, pp. 53-59, 2014.

[8] P. Angleov, Anomaly Detection based on Eccentricity Analysis, *In Proc. 2014 IEEE Symp. Series on Comput. Intel., SSCI-2014, Symp. on Evolving and Autonomus Leanring, EALS2014*, Orlando, FL, USA, 8-12 Dec. 2014, pp.1-8, 2014, IEEE Press, ISBN 978-1-4799-4495-8.

[9] L. Zadeh, Fuzzy Sets, *Information and Control*, vol. 8 (3), 338-353, 1965.

[10] P. Angelov, R. Hyde, DDCAR, *2014 Evolving and Adaptive Learning Systems within 2014 IEEE SSCI*, Florida, USA, 9-12 Dec. 2014, pp….

[11] D. Kangin, P. Angelov, New Autonomously Evolving Classifier TEDA Class, *IEEE International Joint Conference on Neural Networks, IJCNN-2015*, Kilkarne, Republic of Ireland, June 2015, to appear

[12] 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.

[13] P. Angelov, *Autonomous Learning Systems from Data Streams to Knowledge inReal Time*. West Sussex, United Kingdom: John Wiley and Sons, Ltd., 2012.

[14] P. Angelov, Machine Learning (Collaborative Systems), USA patent 8250004, granted 21 August 2012; priority date: 1 Nov. 2006; international filing date 23 Oct. 2007.

[15] P. Angelov, "Anomalous system state identification", patent GB1208542.9, priority date 15 may 2012.

[16] D. Filev, O. Georgieva, An Extended Version of the Gustafson-Kessel Algorithm for Evolving Data Stream Clustering, in *Evolving Intelligent Systems: Methodology and Applications (P. Angelov et al. Eds)*, chapter 12, pp.273-300, John Wiley, 2010.

[17] Central England Temperature data from 1772 to date, available online at http://www.metoffice.gov.uk/hadobs/hadcet/data/download.html