Anomaly Detection based on Eccentricity Analysis

Plamen Angelov
Data Science Group, School of Computing & Communications
Lancaster University
Lancaster, LA1 4WA, UK
p.angelov@lancaster.ac.uk

***Abstract*—In this paper, we propose a new eccentricity- based anomaly detection principle and algorithm. It is based on a further development of the recently introduced data analytics framework (TEDA – from typicality and eccentricity data analytics). We compare TEDA with the traditional statistical approach and prove that TEDA is a generalization of it in regards to the well-known "$n\sigma$" analysis (TEDA gives exactly the same result as the traditional "$n\sigma$" analysis but it does not require the restrictive prior assumptions that are made for the traditional approach to be in place). Moreover, it offers a non-parametric, closed form analytical descriptions (models of the data distribution) to be extracted from the real data realizations, not to be pre-assumed. In addition to that, for several types of proximity/similarity measures (such as Euclidean, cosine, Mahalonobis) it can be calculated recursively, thus, computationally very efficiently and is suitable for real time and online algorithms. Building on the per data sample, exact information about the data distribution in a closed analytical form, in this paper we propose a new less conservative and more sensitive condition for anomaly detection. It is quite different from the traditional "$n\sigma$" type conditions. We demonstrate example where traditional conditions would lead to an increased amount of false negatives or false positives in comparison with the proposed condition. The new condition is intuitive and easy to check for arbitrary data distribution and arbitrary small (but not less than 3) amount of data samples/points. Finally, because the anomaly/novelty/change detection is very important and basic data analysis operation which is in the fundament of such higher level tasks as fault detection, drift detection in data streams, clustering, outliers detection, autonomous video analytics, particle physics, etc. we point to some possible applications which will be the domain of future work.**

***Keywords*—TEDA, typicality, eccentricity, data density, anomaly/novelty/change/outliers detection, data streams**

I. INTRODUCTION

Anomaly detection plays a very important role in Data Analysis as a basis for outlier detection in pre-processing [1], faults detection, and in a dynamic aspect, change detection and *shift* and *drift* in the data concept [2]-[3]. It is also an integral component of the clustering process since the members of a cluster are naturally opposite to being anomalous (they are rather routine, normal or typical). Traditionally, anomaly detection is performed by statistical analysis [4] using frequentistic approach to probabilities and making a number of *prior* assumptions, which do not hold in practice [5].

A widely used principle for anomaly detection is using thresholds or so called "$n\sigma$" principle [4]. The rationale of using thresholds is clear but its disadvantages, too; the use of "$n\sigma$" is also a kind of a threshold which is based on one of the following:

a) assuming a normally distributed random variable and a representatively large amount of data samples; then based on the statistical properties of the Gaussian which guarantee that the vast majority of the data (*>99.7%* if use 3σ) will be considered "normal" and the probability for a data sample/point to be abnormal is *<0.3%*; if use 2σ, for example, then the percentages change to *95.45* and *4.55,* respectively but the principle remains the same;

b) for any distribution (but, still assuming a representatively large amount of independent data samples) one can use so called Chebyshev inequality [6] which states that no more than $1/n^2$ of the data samples/points are more than $n\sigma$ away from the mean (where σ denotes the standard deviation). This means that, the probability to have a data point distant form the mean more than, for example, 3σ is *<1/9* (or *~11%*), for 2σ is *<1/4* (or *25%*). Obviously, these are significantly large amounts of data to be declared anomalous to avoid creating too many false positives. Therefore, they often use in practice for unknown distributions 6σ or even higher n (to guarantee that *<1/36* (or *~3%*) of the data are declared anomalous).

Obviously, such an approach suffers from either requiring strict *prior* assumptions (which rarely materialize in practice) or relaxing too much the condition to avoid false positives to the level where it misses many true positives. Even the 3σ rule (which to avoid excessively large amount of false positives would require the strict assumption of Gaussian/normal distribution) can miss some obvious outliers as it will be illustrated in a very simple example later. In addition, both of these principles (with or without assumption of the type of the distribution) suffer from other disadvantages. For example, the need for a representatively large amount of data samples/points. Another disadvantage is the comparison of a single data point/sample with the average of *all*, instead of comparing pairs of data samples/points; in this way, the information is blurred and is not point-wise and local anymore.

978-1-4799-4494-1/14/$31.00 ©2014 IEEE

In this paper we propose a new, very simple and intuitive condition for anomalies detection which can also be very useful for definition of clusters and other problems like segmentation in images, etc. It does not require any *prior* assumption. It can be formulated as a "$\boldsymbol{\sigma gap}$" that has to exist between the eccentricities of data samples with the larger eccentricity. It can be formulated for a point-wise analysis as well as for the group of points which are in the ε locality/vicinity (where ε is formulated in terms of normalized eccentricity, not value of the variable of interest). While the traditional Chebyshev inequality is too conservative and works "on average" and the newly introduced "$\boldsymbol{\sigma}$ ***gap***" is point-wise but is also conservative for a large/huge number of points (when the points are not sparse with gaps between them) the "ε vicinity" applies locally in regards to the most eccentric data point. This is further detailed in section IV.

The rest of the paper is organized as follows: section II presents a brief introduction of the basic concepts of the TEDA; section III compares TEDA with the traditional "$n\boldsymbol{\sigma}$" statistical analysis; section IV introduces the proposed new anomaly detection principle, "$\boldsymbol{\sigma}$ ***gap***" and provides simple example and illustration; section V outlines how the newly proposed anomaly detection condition within TEDA can be used in other higher level data analytics tasks, and, finally, section VI concludes the paper with directions of the future work.

II. BASIC CONCEPTS OF TEDA

The new data analytics framework based on the typicality and eccentricity (TEDA) was recently introduced [7] aiming to generalize and avoid the well known, but very restrictive assumptions on which the traditional statistical approach and probability theory are based upon, namely:

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

Indeed, probability theory has been developed with *purely* random processes in mind which are used as examples in any textbook, such as games and gambling. It is perfectly suitable for describing such *purely* random processes and variables. However, when we have at hand *real* processes (e.g. climate, economic, physical, biological, social, psychological, etc.) which are **not** *purely* random we do have inter-sample dependence, not normal/Gaussian distributions and not infinite number of observations. That is, the above mentioned assumptions are being violated or ignored. This is a well known problem which is addressed usually by: i) mixture of normal distributions [8]; ii) orthogonalisations such as PCA etc. which are usually offline and require all data to be known in advance and lead to new not real but derived features/input variables. Other alternatives include methods such as particle filtering [22] and information theoretic learning [20]-[21] which are non-parametric, however, they still assume a Gaussian or other kernel to represent the vicinity around the data sample (the field). TEDA, on the contrary does not require any kernel assumption.

The issue with the lack of infinite amount of data and the fact that theoretically it is a requirement is usually being completely ignored.

To avoid all these problems the new TEDA approach [7] was recently introduced which can be characterised as follows:

- ✓ It is entirely based on the data and their mutual distribution in the data space;
- ✓ No prior assumptions are made;
- ✓ No kernels are required;
- ✓ No user- or problem-specific thresholds and parameters are required to be pre-specified;
- ✓ It does not require independence of the individual data samples (observations) – on the contrary, TEDA builds upon (makes use of) their mutual dependence.
- ✓ It also does not require infinite number of observations and can work with as little as *3* data samples.

TEDA can be seen as an alternative statistical framework which can work efficiently with any data **except *pure* random** processes in which individual data samples (observations) are completely independent from each other. For such cases (e.g. gambling, games, white noise, etc.) the traditional probability theory is the best tool to be used since it was developed with such processes in mind and matured over three or more centuries. However, for real data processes - which are the majority of the cases in the Nature - TEDA is more suitable, because it does not rely on assumptions which are not satisfied by such processes. Example of very simple one dimensional (*1D*) real problem of climate analysis will be given for which the traditional statistical analysis fails to identify the anomaly while TEDA and the newly introduced in this paper **"$\boldsymbol{\sigma}$*gap*"** principle is very efficient.

TEDA is based on several new quantities which are all based on the proximity/similarity analysis in the data space. These are not exactly the same as the density used in statistics and other areas, nor the same as entropy. The term *typicality* used is TEDA is somewhat similar to the recently introduced term with the same name in [9] to describe "the extent to which objects are 'good examples' of a concept". By differ from [9] were only conceptual, philosophical considerations are made; TEDA lays the ground of a systematic mathematical framework which is taken further in this paper.

Let us have a data space $\boldsymbol{\mathfrak{X}} \in \mathbb{R}^{\boldsymbol{n}}$, which consisit of n-dimensional data samples/observations/measurements. For this space, we can define a distance $\boldsymbol{d(x,y)}$, e.g. Euclidean, Mahalanobis, cosine, $\boldsymbol{L_1}$, or any other. Then, let us consider the data samples 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 timeinstant when the data sample has arrived. For this reason, index $\boldsymbol{k}$ will be referred as time instant further for simplicity.

TEDA introduces the following characteristics valid for each data sample:

a) *accumulated proximity,* π from a particular point $x \in \mathfrak{X}$, to all remaining data samples up to the k^{th} from a given, j^{th} ($j>1$) data sample calculated when k ($k>1$) data samples are available [7]:

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

where d_{ij} denotes a distance measure between data samples, x_i and x_j, for example Euclidean, Mahalonobis, cosine, etc.

b) *eccentricity* of a particular j^{th} data sample calculated when k ($k>2$) data samples are available (and they are not all the same by value) [7]:

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

c) *typicality* of the j^{th} ($j>1$) sample calculated when k ($k>2$) non-identical data samples are available:

$$\tau_{jk} = 1 - \xi_{jk} \quad k > 2 \quad \sum_{i=1}^{k} \pi_{ik} > 0 \tag{3}$$

It is easy to check that both *eccentricity*, ξ and *typicality*, τ are bounded between *0* and *1* and their normalized counterparts, $\zeta=\xi/2$ and $t = \frac{\tau}{k-2}$ sum up to *1* [7]. They can also be defined both locally and globally same as the accumulated proximity, π. Typicality resembles fuzzy membership functions and the normalized typicality – pdf, however, they differ as typicality does not require the *prior* assumptions that are a must for the probability theory. It represents both the spatial distribution pattern and the frequency of occurrence of a data sample simultaneously and per data sample not "on average" [7].

Moreover, these quantities are applicable online and can be calculated recursively for certain type of distances (such as Euclidean [7,10], Mahalonobis [11,12]). For example, if use Euclidean distance we have):

$$\pi_{jk} = k\left(\| x_j - \mu_k \|^2 + X_k - \| \mu_k \|^2\right) \tag{4}$$

$$\mu_k = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k \quad \mu_1 = x_1 \tag{5}$$

$$X_k = \frac{k-1}{k}X_{k-1} + \frac{1}{k}\| x_k \|^2 \quad X_1 = \| x_1 \|^2 \tag{6}$$

$$\sum_{i=1}^{k}\pi_{ik} = \sum_{i=1}^{k-1}\pi_{i(k-1)} + 2\pi_{kk} \quad \pi_{11} = 0 \tag{7}$$

where μ - recursively updated (local or global) mean;

X is the recursively updated scalar product.

If we use cosine distance normalized to be within the range *[0;1]*, we come to the following expressions:

$$\pi_j = \sum_{i=1}^{k}\frac{1}{2}\left(1 + \frac{x_i x_j}{\|x_i\|\|x_j\|}\right) = \frac{k}{2}\left(1 + \bar{\mu}_k \frac{x_j}{\|x_j\|}\right); \quad \|x\| \neq 0 \tag{8}$$

where

$$\bar{\mu}_k = \frac{1}{k}\sum_{i=1}^{k}\frac{x_i}{\|x_i\|} \quad \|x\| \neq 0 \tag{9}$$

or recursively

$$\bar{\mu}_k = \frac{k-1}{k}\bar{\mu}_{k-1} + \frac{1}{k}\frac{x_k}{\|x_k\|} \quad \|x\| \neq 0 \tag{10}$$

$$\sum_{i=1}^{k}\pi_{ik} = \sum_{i=1}^{k}\frac{1}{2}\left(1 + k\bar{\mu}_k \frac{x_j}{\|x_j\|}\right) = \frac{k}{2}\left(1 + k\bar{\mu}_k\bar{\mu}_k\right); \quad \|x\| \neq 0 \tag{11}$$

Finally, the eccentricity can be determined by:

$$\xi_{jk} = \frac{1 + \bar{\mu}_k \frac{x_j}{\|x_j\|}}{1 + k\bar{\mu}_k\bar{\mu}_k}; \quad \|x\| \neq 0 \tag{11}$$

It has to be stressed, however, that the cosine distance expression is not very informative (useful) for simple 1D cases (as in the illustrative example given), because then $\frac{x_j}{\|x_j\|}$ reduces to $sign(x_j)$ while in the higher dimensional spaces cosine distance provides information about the direction (not just the sign) and the meaning of $\bar{\mu}_k$ is of the average direction of the data vectors in the data space.

Typicality can also be seen as an analogue to the histograms of distributions, but it is in a closed analytical form and does take into account the mutual influence of the neighbouring data samples/observations. It also does not require a large number of data samples (as the histogram does) to be build.

III. A COMPARISON OF THE TEDA-BASED APPROACH WITH THE TRADITIONAL APPROACH OF "$n\sigma$"

It is very interesting to compare the data analysis within TEDA with the traditional probability theory having in mind that ***they are based on different sets of assumptions***. TEDA does not require any *prior* assumptions except that the process that is being observed is not a *pure* random process, but a real (albeit complex) one. Traditional probability theory, on the contrary relies theoretically on a number of strong assumptions which are usually not satisfied in practice and being ignored as described in section I. However, so called "$n\sigma$" analysis for anomalies can be proven to be applicable in both, TEDA and statistical analysis. Moreover, it can be

proven that the results can be ***exactly*** the same (bar the lack of need for any assumptions in TEDA) and simply translate to a condition that we can formulate as ***"(n²+1)/2k"***; where *k>2* is the number of data samples, which concerns the normalized eccentricity. For example, σ translates to *1/k*; 3σ - to *5/k*; 6σ - to *37/2k,* etc. It has to be stressed that this has to be compared with the Chebyshev inequality rather than with the Gaussian because TEDA does not require any assumption in regards to the distribution of the data.

This can easily be proven starting from the definition of the eccentricity in TEDA (using Euclidean distance and the variance definition, $\sigma_k^2 = \sum_{i=1}^{k} \frac{(x_i-\mu_k)^T(x_i-\mu_k)}{k}$ which can be calculated recursively [5]):

$$\xi_k = 2\frac{\sum_{i=1}^{k} d(x_i,x_k)}{\sum_{l=1}^{k}\sum_{i=1}^{k} d(x_i,x_l)} = 2\frac{\sum_{i=1}^{k}\left(x_i^T x_i - 2x_k^T x_i + x_k^T x_i\right)}{\sum_{l=1}^{k}\sum_{i=1}^{k}\left(x_i^T x_i - 2x_l^T x_i + x_l^T x_i\right)} =$$

$$= \frac{2k\left(\mu_k - 2(\mu_k)^T x_k + (x_k)^T x_k\right)}{2k^2\left(2\mu_k - 2(\mu_k)^T \mu_k\right)} = \frac{\mu_k - 2(\mu_k)^T x_k + (x_k)^T x_k}{k\sigma_k^2} =$$

$$= \frac{\sigma_k + (\mu_k)^T \mu_k - 2(\mu_k)^T x_k + (x_k)^T x_k}{k\sigma_k^2} = \frac{\sigma_k + (\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2} =$$

$$= \frac{1}{k} + \frac{(\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2}$$

Therefore,

$$\xi_k = \frac{1}{k} + \frac{(\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2} \quad (12)$$

In fact, for Big Data problems when the value of *k* may become huge and, respectively, equations (4) and (7) can potentially lead to computational problems (hardware dependent, not theoretically restrictive). For such cases, and in general, one can use equation (12) in relation with equation (5) and update of the standard deviation which is similar to (5) and is well known (in general, there is a biased and unbiased form of this update, we use the unbiased one here [10]):

$$\sigma_k^2 = \frac{k-1}{k}\sigma_{k-1}^2 + \frac{1}{k-1}\left\|x_k - \mu_k\right\|^2 \quad \sigma_1^2 = 0 \quad (13)$$

The typicality, can be represented as:

$$\tau(x_k) = 1 - \xi(x_k) = \frac{k-1}{k} - \frac{(x_k-\mu_k)^T(x_k-\mu_k)}{k\sigma_k^2} =$$

$$= 1 - \frac{1}{k}\left(1 + \frac{(x_k-\mu_k)^T(x_k-\mu_k)}{\sigma_k^2}\right) \quad (14)$$

The quantity δ_k,

$$\delta_k = \frac{(x_k-\mu_k)^T(x_k-\mu_k)}{\sigma_k^2} = \frac{k(x_k-\mu_k)^T(x_k-\mu_k)}{\sum_{i=1}^{k}(x_i-\mu_k)^T(x_k-\mu_i)}$$ which can be called *normlaised deviation from the mean*, is very interesting and similar to the negative factor in the exponential of the Gaussian and other kernels (in the Gaussian it is negative and divided by 2), but it has to be stressed that in TEDA **no** *prior* assumptions are made neither for the type of the data distribution, nor for their independence, quantity or existence of any kernel, etc.

Similarly, we can define the normalised typicality and eccentricity:

$$t(x_k) \equiv \frac{\tau_k}{k-2} = \frac{1}{k-2}\left(1 - \frac{1}{k}\left(1 + \frac{(x_k-\mu_k)^T(x_k-\mu_k)}{\sigma_k^2}\right)\right) \quad (15)$$

$$\varsigma(x_k) \equiv \frac{\xi_k}{2} = \frac{1}{2k}\left(1 + \frac{(x_k-\mu_k)^T(x_k-\mu_k)}{\sigma_k^2}\right) \quad (16)$$

Now, we can analyse the Chebyshev inequality in the framework of TEDA in terms of normalized eccentricity. The Chebyshev inequality describes the probability that certain data sample, *x* is more than $n\sigma$ distance away from the mean [6]:

$$(x_k-\mu_k)^2 > n^2\sigma_k^2 \quad (17)$$

$$\frac{(x_k-\mu_k)^2}{2k\sigma_k^2} > \frac{n^2}{2k}$$

$$\frac{(x_k-\mu_k)^2}{2k\sigma_k^2} + \frac{1}{2k} > \frac{n^2}{2k} + \frac{1}{2k}$$

$$\zeta_k = \frac{(x_k-\mu_k)^2}{2k\sigma_k^2} + \frac{1}{2k} > \frac{n^2+1}{2k}$$

So, finally, the condition that provides exactly the same result (but without making any assumptions on the amount of data, their independence etc.) as the Chebyshev inequality can be given as:

$$\zeta_k > \frac{n^2+1}{2k} \quad (18)$$

Moreover, this condition can be expressed for other type of distances also, such as Mahalonobis [12], cosine, etc.

The importance of this result cannot be overstated. It demonstrates that for an arbitrary distribution and as little as three data samples one can analyse per data sample normalised eccentricity; if it exceeds $\frac{n^2+1}{2k}$ then we have ***exactly*** the same result as the Chebyshev inequality would provide.

This can be illustrated on the same simple *1D* example of real data used in [7]: *x={20.2;3;6.4;11.6;8.2;2.2;11.2;5.2;6.2;0.2;1.0;4.8;2.4; 3.8}* which represent the precipitation (rainfall) measured (in *mm*) at Filton station near Bristol, UK in the first two weeks of January 2014 [14], see also Figure 1.

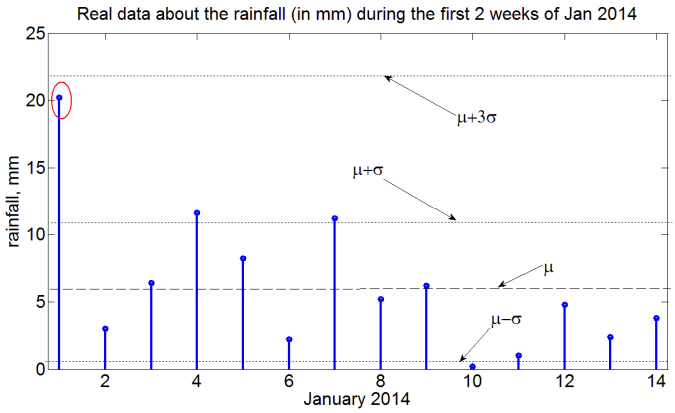

Figure 1 Real rainfall data from Bristol, UK, first two weeks of January, 2014 [7,14].

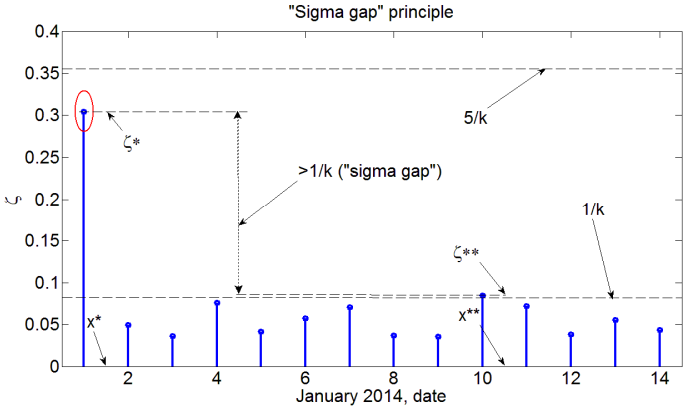

Figure 2 The ***"σgap"*** principle is illustrated on the simple *1D* rainfall data from the first couple of weeks in South-West UK.

It is clearly seen from the plots that the high amount of rainfall (over *20mm*) on the New Year's Day is eccentric, an outlier. However, the traditional (Gaussian or Chebyshev) 3σ or 6σ approach would not detect and identify as an outlier. The most typical amount of rainfall for these two weeks of January 2014 was $\approx 6.2mm$ with $\sigma \approx 5.3mm$. For this problem (data set) $1/k=1/14\approx 0.07143$ in terms of a difference between normalized eccentricities.

IV. The Proposed "σ Gap" Principle for Anomaly Detection Based on Eccentricity

1) The "σgap" principle

Eccentricity can be very useful for anomaly detection, image processing, fault detection, particle physics, etc. It allows per data sample analysis (which can also be done in real time for data streams). In Figure 3 a multidimensional (*4D*) well known and widely used as a benchmark data set about classification of different types of wine [15] is illustrated in terms of the normalized eccentricity per input variable, namely "alcohol", "malic acid", "ash", "alkalinity of ash".

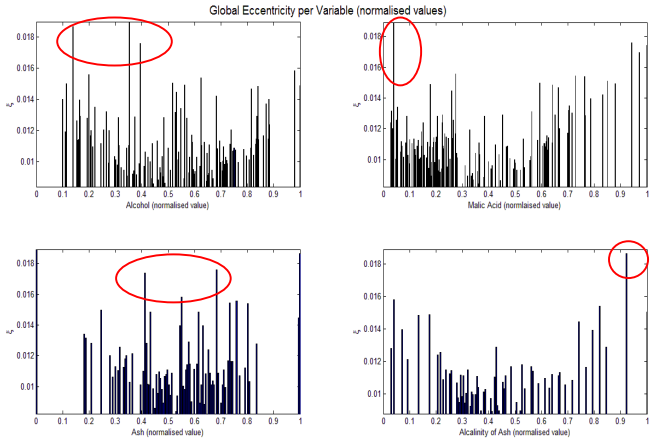

Figure 3 Normalized eccentricity per variable of *4D* data about types of wine from a well-known benchmark problem [13]. No "σ gap" observed – therefore, no outliers detected. However, still some data samples are somewhat standing out (circled), but not declared outliers – this method allows to analyze deeply the data on an individual basis, not "on average". Using the "ε vicinity" around the most eccentric data point multiple outliers can be detected as detailed in the next Figure.

One option is to assume Gaussian distribution of the data (which is not the case in reality) and rely on the 3σ principle. Another, more realistic alternative is to recognize the non-Gaussian character of the data and try to find a suitable mixture of Gaussians. Then,

i) the number of Gaussians needs to be determined, and
ii) it will still not be an exact representation.

Alternatively, one can apply Chebyshev inequality and, for example, the 6σ principle. This, however, will blur the analysis and hide many true positives (real outliers) as it is very conservative. On the other hand, reducing to 3 or 2σ in order to minimize the false negatives (misses of real outliers) would expose too many false positives (noise).

In this paper we introduce so called ***"σgap"*** principle for anomaly detection. The rationale is as follows. The outlier is a data point/sample that stands out and is different from other data samples. In the traditional ***"nσ"*** analysis each sample is compared with the mean/average which is representative of *all* data samples. We propose to analyze the normalized eccentricity of the data samples instead of their values. In this way, the spatial proximity to all other data samples is taken into account by definition and we can compare pairs of data points and in an accumulated and aggregated form with all other points (through the eccentricity). If for a pair of data points starting from the point with the highest normalized eccentricity, $x^1=\{x|\zeta(x^1)=max(\zeta(x_i))\}$ there is a ***"σ gap"*** in terms of their eccentricities than it is an outlier. If denote the

point with the second maximum eccentricity by x^2 than we can define the $\Delta\zeta$ as:

$$\Delta\zeta^{1,2} = \zeta(x^1) - \zeta(x^2) \quad (19)$$

Please, note that upper index 2 means second point in a descending ranking from the maximum and not a square. Same applies further for x^2.

The ***"σ gap"*** condition is very intuitive and is defined as follows:

IF ($\Delta\zeta^{1,2} > n/k$) THEN (x^1 is an outlier) (20)

Indeed, starting from equation (12) we can easily get:

$$\Delta\zeta^{1,2} = 2\left(\frac{\sum_{i=1}^{k} d(x_i, x^1) - \sum_{i=1}^{k} d(x_i, x^2)}{\sum_{l=1}^{k}\sum_{i=1}^{k} d(x_i, x_l)}\right) = \frac{\left(\mu_k^1 - x^1\right)^T\left(\mu_k^1 - x^1\right) - \left(\mu_k^2 - x^2\right)^T\left(\mu_k^2 - x^2\right)}{k\sigma_k^2} \quad (21)$$

Now, combining equations (18) and (21) one can easily get:

$$\left(x^1 - \mu_k^1\right)^T\left(x^1 - \mu_k^1\right) - \left(x^2 - \mu_k^2\right)^T\left(x^2 - \mu_k^2\right) > n\sigma_k^2 \quad (22)$$

This explains the name***"σ gap"***. For example, for $n=1$ we have one variance difference as illustrated in Figure 2. The meaning is that in traditional statistical analysis a point is being compared with the mean while in TEDA we compare points with other points.

It should be stressed that the ***"σ gap"*** analysis has to be done starting form x^1 and moving towards points with smaller eccentricity because the anomalies will naturally be points with higher eccentricity. That means, if the condition (19)-(20) is not satisfied then one can check the same but now for the points with the second and third biggest eccentricities, namely x^2 and x^3, etc.:

$$\Delta\zeta^{2,3} = \zeta(x^2) - \zeta(x^3) \quad (23)$$

IF ($\Delta\zeta^{2,3} > n/k$) THEN (x^3 and x^2 are outliers) (24)

This type of check may go further until a "σ gap" condition is satisfied. Obviously, if (19)-(20) are not satisfied but one of the following similar conditions (such as (23)-(24), for example) there will be more than one (in the latter case – two) outlier points. These outlier points may be close to each other.

2) The "ε vicinity" principle

For such cases - which are more likely when dealing with big data – one can analyze (this can be done online and in a recursive manner) the "ε vicinity" in terms of normalized eccentricity of the most eccentric data point, x^1. The idea is as follows: an ε vicinity ($0 < \varepsilon < 1/k$) around the most eccentric point (x^1) is considered in terms of the normalized eccentricities of these points, not their nominal values, x:

$$x^\varepsilon = \{x_i | \ \zeta(x^1) - \zeta(x_i) < \varepsilon\} \quad (25)$$

In this way, all the data points are split in two groups – one is the set of potentially eccentric points defined by equation (25) and all remaining data points. Then, we can calculate the mean of the normalized eccentricities of all potentially eccentric points defined by equation (25), $\overline{\zeta}^\varepsilon$. We can also easily find the mean normalized eccentricity of all remaining points, $\overline{\zeta}^t$ where t stands for typical. All potentially eccentric points will be declared outliers if a condition similar to equation (20) is satisfied:

$$\overline{\zeta}^\varepsilon - \overline{\zeta}^t > n/k \quad (26)$$

In Figure 4 we used as a basis exactly the same real (quite simple, *1D* low volume) data about the rainfall in South-West UK in January 2014 with the only difference that we replaced the true data sample for 14 January 2014 ($x=3.8mm$) with an artificially high value of $x=19.5mm$ simply to demonstrate the effect of ε vicinity when there are more than a single point forming the ***"σ gap"***. The data is still extremely simple and didactic in nature. For this set of data $\overline{\zeta}^\varepsilon = 0.1871; \overline{\zeta}^t = 0.0521$.

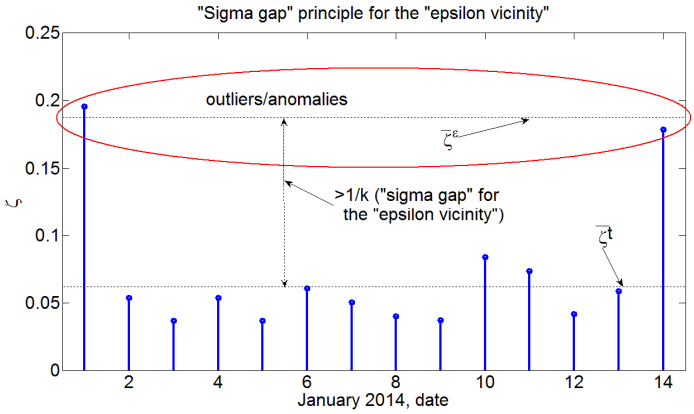

Figure 4 The ***"σ gap"*** principle for an ε vicinity around the most eccentric data point. The bulk of the data samples is real (13 out of 14 points) and same as in Figure 2 with the only artificial introduction of the last data sample for 14 January in close vicinity of the most eccentric data sample (1 January 2014). Both points that form the ***"ε vicinity"*** in this simplified example are identified as anomalies/outliers despite being within 3σ without any prior assumption on the data distribution which is clearly not Gaussian. This is based on the ***"σ gap"*** between the means of the eccentric and typical points which is also clearly visible in this didactic *1D* example. In addition, the ***"ε vicinity"*** condition helps distinguish between anomaly (technical fault, novel object detected in video, etc.) and noise. Noise will express itself through small number of points scattered randomly while a "meaningful" anomaly would invoke a number of abnormal points in an ***"ε vicinity"***.

The following simple algorithm can be proposed for identifying anomalies:

1. Calculate normalized eccentricity of a point.

2. Keep the point with the maximum normalized eccentricity, x^1 second maximum normalized eccentricity, x^2, etc.
3. Check the ***"σ gap"*** condition (19)-(20)
4. If it is satisfied, declare the point x^1 an outlier; end.
5. Else, check condition (23)-(24), etc. OR for certain ε check condition (24) in relation to (23).
6. If (21)-(22) is satisfied declare both x^1 and x^2 outliers; If (25) is satisfied declare all points form the ***"ε vicinity"*** outliers; end.
7. Else continue in a similar manner until outlier is detected or no more data points are available for interrogation.
8. End

Note, such points may be able to form a cluster/data cloud between themselves if they satisfy conditions for forming clusters/data clouds in terms of minimum number of points and density, but still they will be considered as outliers in regards to all other data points.

In the case of "ε vicinity" the "σ gap" condition translates to "if the difference between the mean of all points (global mean) and the mean of the subset of eccentric points on one hand and the global mean and the mean of the subset of typical points on the other is greater than $n\sigma$ then all eccentric points from the ε vicinity are outliers.

An illustration of an anomalous point (which the traditional 3σ (Gaussian, chi-square, etc.) or even more 6σ (Chebyshev) type analysis is unable to pick up) is shown in Figure 2 for the real (though simple, *1D*) data set. Applying the ***"σ gap"*** principle, it can easily be seen that the point is identified as an outlier, which is also obvious from Figure 1 that is a logical conclusion. Similarly, but for slightly manipulated data (where only a single data point out of 14 which represents the rainfall on 14 January 2014 near Bristol, UK) a group of two points is easily identified as outliers/anomalies although neither 3 nor 6σ type analysis was able to pick this up). Moreover, no prior assumptions about Gaussian or any other specific type of distribution of the data is made nor about their number (which is indeed very low in this case, but can be as big as necessary), nor about the orthogonality of the individual data samples.

The proposed ***"σ gap"*** principle combines the accumulated proximity information to ***all*** data samples/points with the pair-wise comparison between two specific suspected points. The traditional statistical *3* or 6σ analysis considers only ***"on average" proximity*** and not the distance between a candidate for outlier data point and the closest to it data point, for example. Therefore, it depends heavily on the distribution of the points which is usually not known but assumed to be a standard/simplified one or is very conservative to the extend to be impractical (Chebyshev condition). The proposed ***"σ gap"*** principle uses *local spatial distribution information* about the neighbourhood of the suspected point while the traditional statistical analysis ***ignores*** such local information. The latter is also based on ***frequency of occurrence*** of data samples rather than on their ***mutual spatial proximity*** [7]. As a result, the newly proposed approach is able to identify outliers which are obvious by manual visual inspection as demonstrated with the simple example of two weeks of rainfall in South-West England, but were unpicked by the traditional analysis.

V. Potential Applications of the Proposed "σ gap" Principle for Higher Level Data Analytics

TEDA and the newly proposed "σ gap" principle can be very useful for development of new clustering, data clouds formation [16], classification, multi-model prognostic, multi-model controllers, machine health monitoring and prognostics, self-calibrating inferential sensors and the methods and applications.

It is well known that clustering is pivotal for pattern recognition, data mining, machine learning, etc. Forming data clouds is a similar process introduced in [16]. Forming clusters, and respectively data clouds, is traditionally based on different principles (it is predefined number in k means approach, depends on a threshold in the hierarchical clustering, is based on potential or density [5,10] or spatial proximity in the data space, etc.). The newly introduced "σ gap" principle can be used as a trigger to form or stop forming new data clouds or clusters. Traditionally they use the distance between a point and the mean (the norm) but, as it was pointed earlier, in regards to the anomaly detection it only carries aggregated information "on average" and blurs the local vicinity information. The newly proposed "σ gap" principle, on the other hand, does combine the local and global information. This will be exploited in development of new clustering and data cloud formation (data clouding) methods [17].

Another important aspect of the TEDA and the newly proposed "σ gap" principle is their ability to be applied successfully to Big Data and data streams. The former is related to the nature of the algorithm to be parallelizable and recursive (the means and other accumulated quantities are sums which themselves can be represented as partial sums and the algorithm can easily be broken down into *N* machines/cores/GPUs etc. for parallel execution rendering exactly the same result but significantly faster [18]). The latter is related to the recursive and one pass (non-iterative) nature of algorithms [10]. Further exploration of TEDA and the newly introduced "σ gap" principle to Big Data and data streams and to specific application areas (e.g., but not limited to, real-time video analytics, climate data analysis [17], autonomous fault detection and diagnostics, etc.) will be a matter of further imminent publications.

VI. Conclusions

In this paper, the recently introduced data analytics framework TEDA is further developed by the introduction of the so called "σ gap" principle for individual data points and for the and "ε vicinity" of the most eccentric data point and by a thorough comparison with the traditional statistical approach and proving that TEDA is a generalization of it in terms of the "$n\sigma$" analysis of outliers. It has been theoretically proven that TEDA provides exactly the same result as the "$n\sigma$" analysis of

outliers, however, without the restrictive requirement for *prior* assumptions that are made for the traditional approach to be in place. Moreover, it offers a non-parametric, closed form analytical descriptions (models of the data distribution) to be extracted from the real data realizations, not to be pre-assumed. In addition to that, for several types of proximity/similarity measures (such as Euclidean, cosine, Mahalonobis) it can be calculated recursively, thus, computationally very efficiently and suitable for real time and online algorithms. In addition to that, TEDA offers per data point analysis which combines the aggregated accumulated spatial proximity information with the local per data sample and ε vicinity analysis. It provides a single scalar measure (within the range *[0;1]* or *[0;100%]* to represent the typicality-eccentricity of a multidimensional vector data set/stream.

The newly introduced ***"ε vicinity"*** condition helps distinguish between anomaly (technical fault, novel object detected in video, etc.) and noise. Noise will express itself through small number of points scattered randomly while a "meaningful" anomaly would invoke a number of abnormal points in an ***"ε vicinity"***.

Further, we demonstrated on very simple and intuitive real data from the rainfall in South West England that an obvious outlier would have been left undetected by the traditional statistical analysis while the newly proposed "σ gap" analysis principle is effective in picking it up. Furthermore, this new principle can be very useful in development of more efficient and effective new clustering, data clouding, classification and other higher level data analytics approaches which will be a matter of forthcoming publications. In terms of data streams, it can be useful for detecting shifts and drifts in the data concept.

VII. ACKNOWLEDGEMENTS

The author would like to thank Jose Principe, Dmitry Kangin, Denis Kolev, Bruno Costa for the useful discussions on TEDA.

REFERENCES

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM Computing Surveys (CSUR)*, Vol. 41 (3), article No.15, July 2009, ACM, NY, USA, DOI: 10.1145/1541880.1541882

[2] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine Learning,* vol. 23, pp. 69-101, 1996.

[3] J. Gama, I. Žliobaité, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys,* vol. 1, p. 35, 2013.

[4] A. Bernieri, G., Betta, C., Liguori, "On-line fault detection and diagnosis obtained by implementing neural algorithms on a digital signal processor". *IEEE Transactions on Instrumentation and Measurement*, vol. 45, 894–899, 1996.

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

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

[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] 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.

[9] D. Osherson, E. E. Smith, Discussion: On typicality and vagueness, *Cognition*, v.64: 189-206, 1997.

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

[11] D. Kolev, P. Angelov, G. Markarian, M. Suvorov and S. Lysanov, ARFA: Automated Real-time Flight Data Analysis using Evolving Clustering, Classifiers and Recursive Density Estimation , *Proc. IEEE Symposium Series on Computational Intelligence SSCI-2013*,16-19 April 2013, Singapore, ISBN 978-1-4673-5855-2/13, pp. 91-97.

[12] 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, in preparation

[13] R. Hyde, P. Angelov, Data Density Based Clustering, *14th UK Workshop on Computational Intelligence*, UKCI 2014, Bradford, UK, 8-10 September 2014, to appear.

[14] http://www.martynhicks.co.uk/weather/data.php?page=m01y2014, accessed 6 February 2014

[15] K. Bache and M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

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

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

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

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

[20] J.C. Principe, D. Xu and J. Fisher, In: *Unsupervised Adaptive Filtering* (S. Haykin Ed.), pp 265-319, Wiley, 2000.

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

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