Outside the Box: An Alternative Data Analytics Frame-work

Submitted: 13th February 2013; accepted 26th February 2013

Plamen Angelov

DOI: DOI 10.14313/JAMRIS_2-2014/16

Abstract: *In this paper, an alternative framework for data analytics is proposed which is based on the spatially-aware concepts of eccentricity and typicality which represent the density and proximity in the data space. This approach is statistical, but differs from the traditional probability theory which is frequentist in nature. It also differs from the belief and possibility-based approaches as well as from the deterministic first principles approaches, although it can be seen as deterministic in the sense that it provides exactly the same result for the same data. It also differs from the subjective expert-based approaches such as fuzzy sets. It can be used to detect anomalies, faults, form clusters, classes, predictive models, controllers. The main motivation for introducing the new typicality- and eccentricity-based data analytics (TEDA) is the fact that real processes which are of interest for data analytics, such as climate, economic and financial, electro-mechanical, biological, social and psychological etc., are often complex, uncertain and poorly known, but* ***not*** *purely random. Unlike, purely random processes, such as throwing dices, tossing coins, choosing coloured balls from bowls and other games, real life processes of interest do violate the main assumptions which the traditional probability theory requires. At the same time they are seldom deterministic (more precisely, have always uncertainty/noise component which is non-deterministic), creating expert and belief-based possibilistic models is cumbersome and subjective. Despite this, different groups of researchers and practitioners favour and do use one of the above approaches with probability theory being (perhaps) the most widely used one. The proposed new framework TEDA is a systematic methodology which does not require prior assumptions and can be used for development of a range of methods for anomalies and fault detection, image processing, clustering, classification, prediction, control, filtering, regression, etc. In this paper due to the space limitations, only few illustrative examples are provided aiming proof of concept.*

Keywords: *data density, proximity measures, RDE, data analytics, data-driven approaches, machine learning, Bayesian*

1. Introduction

Probability theory was around for over two centuries [1]. It is well established and widely (over)used. Its basis was set up by Thomas Bayes, generalised later by Pierre-Simon Laplace and other researchers based on observations of *purely* random processes, such as games and gambling. It is perfectly suitable for describing such *purely* random processes and variables. However, it is also (extremely) widely (over)used to describe real world processes which are not *purely* random and have inter-sample dependence, not normal distributions and may have small number of observations. For example, climate, economic, physical, biological, social, psychological and many other real processes are complex and difficult to tackle using first principle or expert-based models. The traditional probability theory, on the other hand, is based on several assumptions which do not hold in practice, such as:

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

The first assumption is fully satisfied for *pure* random processes, but not for real processes which are usually of interest. Therefore, the application of traditional probability theory for *pure* random processes is justified, but the same is not necessarily the case for the real processes which are the vast majority of applications of interest.

In this paper, a new systematic framework for data analytics is proposed which requires no *prior* assumptions or kernels, user- or problem-specific thresholds and parameters to be pre-specified. It is entirely based on the data and their mutual distribution in the data space. It does not require independence of the individual data samples (observations); on the contrary, the proposed approach builds upon their mutual dependence. It also does not require infinite number of observations and can work with as little as *3* data samples.

The new *typicality* and *eccentricity* based data analytics (TEDA) is an alternative statistical framework which can work efficiently with any data except *pure* random processes when individual data samples (observations) are completely independent from each other. For such *pure* random data the traditional probability theory is the best tool to be used. However, for real data processes – which are the majority of the cases – we argue that TEDA is better justified, because it does not rely on assumptions which are not satisfied by such processes.

The term *typicality* was used recently [2] to describe "the extent to which objects are 'good examples' of a concept". By differ from [2] were only conceptual, philosophical considerations are made, in this paper a systematic mathematical framework is introduced.

Eccentricity can be very useful for anomaly detection, image processing, fault detection, etc. Both *typicality* and *eccentricity* can be very useful for development of new clustering, classification, multi-model prognostic, control, soft sensors, etc.

In the remainder of this paper the proposed new TEDA methodology will be described first in section 2. In section 3, some simple examples will be provided mostly aiming proof of concept. Next, in section 4, an anomaly detection approach based on *eccentricity* will be outlined. In section 5 the clustering, classification in section 6 and prediction and control in section 7, all within the TEDA framework are outlined. Finally, section 8 concludes the paper.

2. Description of the proposed methodology

Let us start with data samples (observations) that we may have, $x \, R^n$ (where n is the number of features/characteristics; in Fig. 1 $n=2$ for illustration purposes; in the rest of the paper we will use $n=1$ without any limitations to the concept which is applicable for any positive integer). If we have a single or just two data samples (observations) there is no much sense to introduce the value of its *typicality* and *eccentricity*. It will be the only value observed/recorded or the only single distance between the two samples, $k=2$ (they will be equally untypical except the extreme case when they coincide when they will be equally typical). For any number of data samples, $k>1$ we can define the distance between them, d. This distance/proximity measure can be of any form, e.g. Euclidean, Mahalonobis, cosine, Manhattan/city/L_1, etc. Let us denote the distance between two data samples, x_i and x_j by d_{ij}.

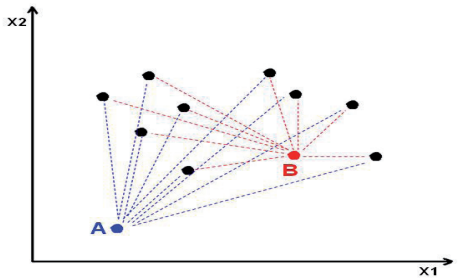

Fig. 1. A 2D data distribution (A is a rather eccentric data point; B – a typical one)

We can also calculate the accumulated proximity/sum distances, π to all available data samples from a given, j^{th} ($j>1$) data sample calculated when k ($k>1$) data samples are available:

$$\pi_k(x_j) = \pi_j^k = \sum_{i=1}^{k} d_j \quad k>1 \quad j>1 \tag{1}$$

The *eccentricity* of a particular j^{th} ($j>1$) data sample calculated when k ($k>2$) data samples are available (and they are not all the same by value) is defined as the relative (normalised) π of that data sample as a fraction of π's of all other data samples:

$$\xi_j^k = \frac{2\pi_j^k}{\sum_{i=1}^{k}\pi_i^k} = \frac{2\sum_{i=1}^{k} d_j}{\sum_{l=1}^{k}\sum_{i=1}^{k} d_l} \quad \sum_{i=1}^{k}\pi_i^k > 0 \quad j>1 \quad k>2 \tag{2}$$

The coefficient 2 is due to the fact that each distance is counted twice and can be seen as a normalisation coefficient.

The *typicality* of the j^{th} ($j>1$) data sample calculated when k ($k>2$) non-identical data samples are available is defined as the complement of the *eccentricity*, ξ of that data sample:

$$\tau_j^k = 1 - \xi_j^k \quad k>2 \quad j>1 \quad \sum_{i=1}^{k}\pi_i^k > 0 \tag{3}$$

It is easy to check that both *eccentricity*, ξ and *typicality*, τ are bounded:

$$0 < \xi_j^k < 1 \quad \sum_{i=1}^{k}\xi_i^k = 2 \quad \sum_{i=1}^{k}\pi_i^k > 0 \quad \forall k>2 \quad \forall j>1 \tag{4a}$$

$$0 < \tau_j^k < 1 \quad \sum_{i=1}^{k}\tau_i^k = k-2 \quad \sum_{i=1}^{k}\pi_i^k > 0 \quad \forall j>1 \tag{4b}$$

These definitions of *eccentricity* and *typicality* resemble fuzzy set membership functions since being values between *0* and 1, summing up to a value larger than 1 (for τ and k=3 it sums up to 1). We can also introduce normalised *eccentricity* and *typicality* which integrate to 1:

$$\zeta_j^k = \frac{\xi_j^k}{2} \quad \sum_{i=1}^{k}\zeta_i^k = 1 \quad 0 < \zeta_i^k < \frac{1}{2} \quad k>2 \quad j>1 \quad \sum_{i=1}^{k}\pi_i^k > 0 \tag{5}$$

$$t_j^k = \frac{\tau_j^k}{k-2} \quad \sum_{i=1}^{k} t_i^k = 1 \quad 0 < t_i^k < \frac{1}{k-2} \quad k>2 \quad j>1 \quad \sum_{i=1}^{k}\pi_i^k > 0 \tag{6}$$

Normalised *eccentricity* and *typicality* resemble probability distribution function (pdf) in that they sum to 1, but they are different as they do not require the *prior* assumptions that are a must for the probability theory and they represent both the spatial distribution pattern and the frequency of occurrence of a data sample.

All of the above definitions are applicable to data streams (online, when *k* is incrementally increased). In case of data sets (offline, fixed amount of data, *k*) the upper index, *k* can be omitted in all notations because it only indicates based on how many data samples the respective value has been calculated. The above definitions are *global* (defined over all available data); one can also define *local eccentricity* and *typicality.* They can also be very useful for *local* regions, groups/clusters/data clouds, classes (then summation is only over the data samples concerned), see also section 6.

The *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. When normalised *eccentricity* is above *1/k* the data sample is rather untypical/eccentric/anomalous. If the value of *typicality* is above 1*/k* then the data sample is rather typical.

It can also be proven that both *eccentricity* and *typicality* can be calculated recursively by updating only the *global* or *local* mean, μ and scalar product, *X* for the cases when Euclidean square distance [3], [4] is used and similarly if cosine [5] or Mahalonobis square distance [6] are used. For example, for the Euclidean square distance, without limiting the scope of applicability of the *typicality* and *eccentricity* in general, we have [3], [4]:

$$\pi_j^k = k\left(||\, x_j - \mu^k\,||^2 + X^k - ||\,\mu^k\,||^2\right) \tag{7}$$

$$\mu^k = \frac{k-1}{k}\mu^{k-1} + \frac{1}{k}x_k \quad \mu^1 = x_1 \tag{8}$$

$$X^k = \frac{k-1}{k}X^{k-1} + \frac{1}{k}||\, x_k\,||^2 \quad X^1 = ||\, x_1\,||^2 \tag{9}$$

where μ– recursively updated (local or global) mean; *X* is the recursively updated scalar product.

Furthermore, we do not need to calculate

$$\sum_{i=1}^{k}\pi_i^k$$

each time, but we can update it recursively by:

$$\sum_{i=1}^{k}\pi_i^k = \sum_{i=1}^{k-1}\pi_i^{k-1} + 2\pi_k^k \quad \pi_1^1 = 0 \tag{10}$$

The coefficient *2* is due to the fact that each distance counts once from the k^{th} point towards the i^{th} point and once from the i^{th} point towards the k^{th} point. It can be proven [12] and it is obvious from (7) that the minimum value of *eccentricity* (and respectively, the maximum of the *typicality*) is obtained for the data points that are closer to (or coincide with) the mean, μ^k which is quite logical. It has to be stressed that this applies globally as well as locally.

Now, we can formulate the following recursive procedure:

Algorithm 1. TEDA (Euclidean square distance used for d)
Initialise $k=1$; $j=1$; x_1; $X^1 = ||\, x_1\,||^2$; $\mu^1 = x_1$; $\pi^1{}_1 = 0$;
WHILE data points from the data stream are available (or until not interrupted) **DO**

1. Read the next (k:=k+1) data point x_k;
2. Update
   a. μ^k from equation (8);
   b. $\mathbf{X}^k$ from equation (9);
3. For $\forall 1 \le j \le k$ compute π_j^k from equation (7);
4. Update $\sum_{i=1}^{k}\pi_i^k$ from equation (10);
5. For $\forall 1 \le j \le k$ compute:
   a. ξ_j^k from equation (2);
   b. τ_j^k from equation (3);
   c. ζ_j^k from equation (5);
   d. t_j^k from equation (6);

End WHILE

The recalculation/update of the values of *eccentricity* and *typicality* based on *k* data samples from the same values based on *k*–1 data samples (observations) can be seen as similar to the *posterior* probabilities update in the Bayesian rule. The *prior* estimation for feasible but not yet observed data points can be done by interpolating the existing ξ, ζ, τ, *t*. Interpolation can be *local* or *global*, linear or more complex.

3. TEDA Primer

Due to the space and time limitations in this paper only the basic concept will be laid down and several illustrative examples aiming proof of concept. Further publications will detail and expand this new theoretical framework for objective data analytics.

Let us consider an extremely simple data stream which consists of just three data samples (these may be thought of – without limiting the generality of the concept – as values of the temperature in °C):

$$y = \{20;\ 12;\ 10\} \tag{11}$$

Obviously, $k = 3$. We can easily get:

$$\xi^3 = \{0.9;0.5;0.6\} \quad \tau^3 = \{0.1;0.5;0.4\}$$

$$s^3 = \{0.45;0.25;0.3\} \quad t^3 = \tau^3 \tag{12}$$

We can see that the sum of *eccentricity* values is = 2 and of the *typicality* values is =1; they are between 0 and 1 as expected. Similarly, the normalized *eccentricity* and *typicality* both sum up to 1 with the normalized *eccentricity* being in the range [0;0.5] and normalized *typicality*- in the range [0;1] as expected. Moreover, the normalized *typicality* of y_2 is above 1/3, which means it is a *typical* value of the temperature (based on these three observations). We can also

see in Fig.2 top line of plots that the *eccentricity* of y_1=20°C is substantially higher than that of the other data samples and the normalized *eccentricity* (0.45) is *>1/3.* The *typicality* of y_2=12°C is highest; the normalized *typicality* of y_3=10°C is also >1/3, but less obvious than that of y_2=12°C.

The above observations are quite logical and, importantly, we did not made **any prior assumptions** on the number of data points, their distributions, kernels, we did not used any expert or first principles knowledge. Yet, we derived objectively the common knowledge fact that y_1=20°C is rather *eccentric*, unusual (for England), thus, a candidate for anomaly being declared while y_2=12°C is the most typical one, thus, a candidate for a prototype.

If we have an additional observation of 18°C then we can estimate *posterior* values of ξ, ζ, τ, *t* using the procedure described above and the result will be different because it will be based on four data samples observed not three (fact, not *prior* estimation, similar to the probability theory's Bayesian rule), z_4:

$$z^4 = \{y;17\} = \{20,12,10,17\}$$

We can use the procedure (Algorithm 1) to easily get the following:

$$\xi^4 = \{0.6;0.43;0.54;0.43\} \quad \tau^4 = \{0.4;0.57;0.46;0.57\} \tag{13}$$

We can check that both ξ and τ sum up to 2 and ζ and *t* sum up to 1 and that the range of ξ and τ is between 0 and 1 while of ζ and *t* it is between 0 and 1/2.

We can also observe (compare in Fig. 2 the two lines of plots) that by adding just a single point close to the data sample that was eccentric the whole pattern is changing with all four data samples becoming more balanced in terms of their *eccentricity* and *typicality* with higher normalized *typicality* of 0.286 (which is also notably higher than 1/*k*=1/4, but not so prominently now), for the two inner samples, z_2 = 12°C and z_4 = 17°C which is quite logical for these observations and for the UK climate unlike if these were numbers of bingo which would have been completely independent indeed. Note that we do not need to assume any distribution or to parameterize it a priori. We can derive two local distributions around the two data samples that have normalized typicality above 1/*k*, namely z_2 and z_4 but this would be knowledge extracted/learned automatically from the data.

This is quite logical because now we have two very simple groups/data clouds/clusters of data and modes of the distribution. It is important to stress that even if it is not strongly obvious the two modes of the distribution were derived from the data automatically (they both are above 1/*k* = 1/2), not assumed or pre-defined! In TEDA there is no need for prior assumptions. All the useful information is contained in the data distribution.

Finally, let us consider a more realistic data stream with a larger amount of data: v^{14} = {20.2, 3, 6.4, 11.6, 8.2, 2.2, 11.2,5.2, 6.2,0.2, 1, 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 [11]. Due to the larger amount (*k*=14) of (still *1D, n*=1) data we used the procedure described in the Algorithms 1 which is based on the square Euclidean distance and recursive calculations.

It is clearly seen from the plots that the high amount of rainfall (over 20 mm) on the New Years Day is rather untypical. It is also untypical for these first two weeks of January 2014 to have low level of precipitation, close to 0. The most typical amount of rainfall for these two weeks of January 2014 was 6.2 mm.

Even with these extremely simplistic (handcrafted) examples the difference that TEDA brings in comparison with the traditional probability theory, deterministic, possibilistic, fuzzy and other representations is obvious. For example, traditional probability theory would suggest equal (1/3 or 1/4) probability for all samples (we also do not

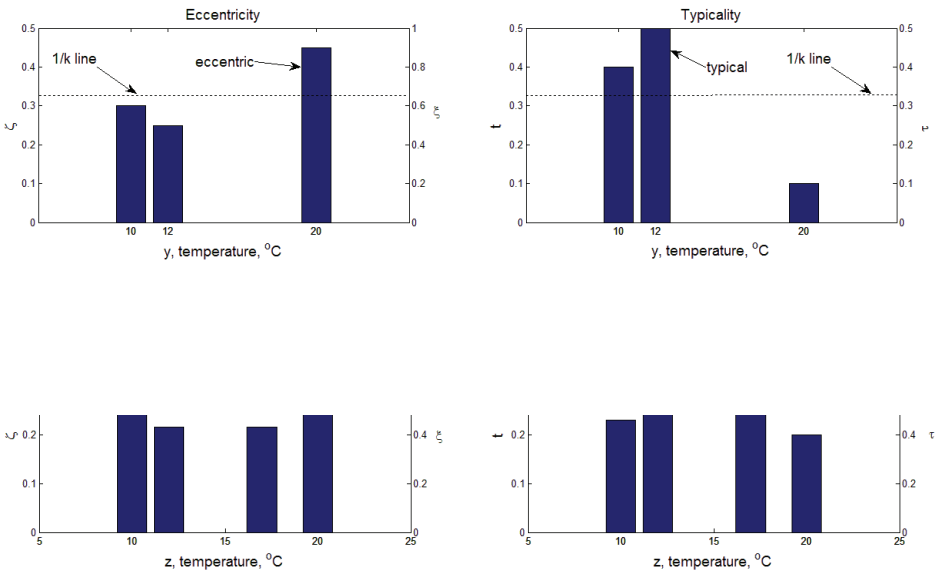

Fig. 2. The left column represents the eccentricity; the right one – typicality; top line of plots corresponds to y and the bottom one – to z. Right hand side vertical axes on each plot represent the normalized eccentricity/ typicality, ζ/t resp.)

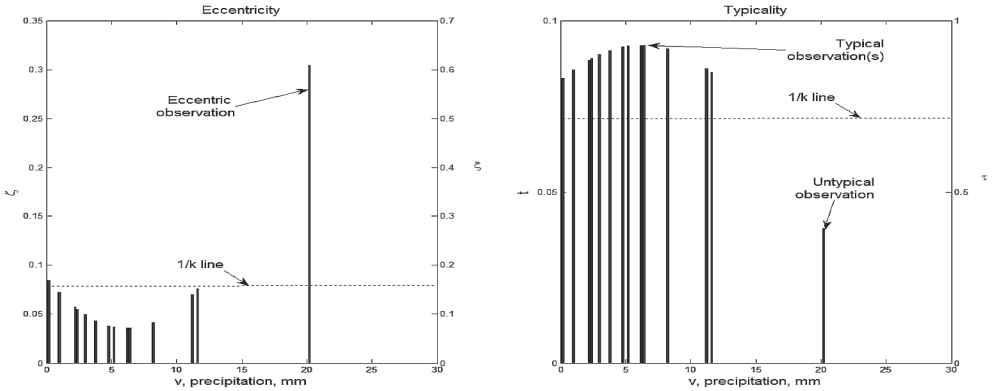

Fig. 3 Real rainfall data from Bristol, UK, first two weeks of January, 2014. The notations are same as in Figure 2

need to build histograms which provide no information about (completely ignore) the inter-sample influence. An alternative which is often (over)used is to impose/assume a distribution or a kernel, for example Gaussian/normal or another type, to determine its parameters (where to position it, how much will be the spread). To escape these problems, sometimes, one can also use a mixture of distributions, but then the parameters that need to be determined are even more and the problem is not fully solved as the distributions are approximations of the real/true ones. On the other hand, in comparison with the fuzzy set theory [7] we do not need to ask experts, to build membership functions. What we only need is to calculate the *eccentricity* and *typicality* of each observation and we get (recursively/computationally efficiently using (7)–(10)) in a closed analytical form (equations (2)–(3)) the distributions.

Moreover, the information that we have in ξ and τ (resp. ζ and t) is closer to the nature of real processes (not the *pure* random ones and not subjective ones, for which the traditional probability and fuzzy set theories, respectively are more appropriate). In particular, from Fig. 2 we can see that y_2 = 12°C is the most typical data sample, but its degree of *typicality* is significantly reduced if another sample is added. On the contrary, the data sample/observation y_1=20°C is rather eccentric initially but becomes neutral (both *typicality* and *eccentricity* are about $1/k$) when we add the fourth sample. Now, if we try to answer the question, "*What is the most typical or likely temperature (or amount of rainfall) based on the observations we have (y and z for the temperature and the real observations of the rainfall, v)".* TEDA suggests that based on *3* observations, *y* the most typical/likely temperature is 12°C ($\tau=t=0.5$ that is 50%). Based on the same limited number of observations we can also conclude that to have a temperature 10°C is also not untypical ($\tau=t=0.4$ that is 40%), but this is now much closer to *1/k=1/3*. To have a temperature 20°C is possible, but not very typical for England ($\tau=t=0.1$ that is 10% based on these three observations). If we have another observation of z_4=17°C, however the *typicality* changes significantly ($t=0.2$ that is 20%) because it is now based on a quite different (balanced) data pattern, but this is still below the 1/k level which means that it is still untypical (but less so based on these four observations in comparison with the three observations only).

For the same observations the traditional probability theory [1] would suggest p=1/3 (same for all the 3 observations) or we would need to choose and parameterize distribution(s). However, the problems of how many distributions to consider, which type of distributions to use for particular data sets, what their parameters are left for the problem solver to decide. Many prefer to approximate the real/true distributions with some smooth functions (such as Gaussian and others), but these are just approximations.

The reason for the difference between TEDA and the traditional probability theory is the spatial awareness which in the traditional probability theory is ignored, but in real processes is a fact. For example, for the very simplistic example we considered above, 2 data samples are quite close and influence each other. TEDA offers instead an automatic mechanism to extract the real/true data distributions and a closed analytical recursive form (which can be differentiated and analyzed) and this is dictated by the data pattern, not pre-defined or assumed. In addition, it does take into account inter-sample influence which is typical for real processes (not *pure* random ones).

4. Anomaly Detection Based on *Eccentricity*

Anomaly detection in TEDA can easily and intuitively be done on the basis of *eccentricity*. For example, any data sample that has high normalised *eccentricity* ($\xi>1/k$) is a suspected anomaly. Different algorithms can be developed and applied to image processing and video analytics, fault detection, user behaviour modelling, etc. One can take into account not only the absolute value of ξ and ζ, but also the context of the problem at hand. However, the *eccentricity* offers new angle of view towards the problems in comparison with the traditionally used probability because of the reasons mentioned above. For example, the *eccentricity* of the data sample/observation y_1=20°C is much higher than that of the other data samples but the probability of all samples is equal (1/3). No distributions or kernels needs to be assumed, no need to have large amount of data, the distribution of *typicality* and *eccentricity* can be extracted from the data in a closed analytical form, (2)–(3), and is exact (not approximated), recursively updated. One emerging area of research and interest for the society is the study of extreme natural and man-made (anthropogenic) events (including, but

not limited to climate, volcanoes, earthquakes, tsunami, nuclear and other disasters, terrorism, etc.) – traditional probability theory is limited in studying the probability of occurrence of such events and this is also limited by the amount of available data, representativeness of the 'training data' which in such problems are a bottleneck, distributions which are not normal etc. TEDA framework offers not only a convenient approach to easily detect anomalies, but also to estimate the degree of severity (how bigger ζ is in comparison with $1/k$ and, respectively, how smaller τ is in comparison with $1/k$).

5. Clustering and Data Clouds Based on *Typicality*

Clustering is an important part of pattern recognition, machine learning, data mining [1] and many other related areas, including autonomous learning systems [3]. The term "data cloud" was introduced in the so called AnYa framework [8] and differs from clusters by the fact that data clouds have no specific shape, parameters, and boundaries. In TEDA, data clouds (or clustering if that is the preferred form of data partitioning) can be formed on the basis of the *typicality*. For example, data sample that has the highest τ is logical to be selected as the focal point/prototype or a centre of the first data cloud (or cluster). There can be different ways to form the other data clouds (or clusters) but their focal points/prototypes (or centres) will also have high *typicality* (e.g. it is logical to require $\tau>1/k$). For example, a zone of influence/radius can be defined and the data points that are outside the zone of influence/radius of the data point with the maximum τ (τ_{max}) and have $\tau>1/k$ can be considered as candidates to be prototypes/focal points of the next data clouds/clusters and the point with the maximum τ but out of these points only (except the data points that fall in the zone of influence associated with the previous focal point(s)) will, be the obvious new focal point, etc. until there are points that satisfy these conditions.

It is important to stress that within TEDA framework one can extract automatically and recursively closed form analytical expressions of the real/true distributions of *local typicality* and *eccentricity* with the former resembling the membership functions or pdf but being conceptually different (we can argue richer because it takes into account objectively both the frequency of occurrence and the spatial distribution and mutual influences). For example, if we have data of two clusters/data clouds, say coloured blue and red, we can automatically and recursively extract from the real data distribution, ξ_{red}, τ_{red}, and ξ_{blue}, τ_{blue}.

In an online and evolving scenario in the memory only the accumulated values per data cloud/cluster can be kept (not for each data sample – see steps 2 and 4 of the TEDA procedure (Algorithm1 in section 2) – these include:

$$\mathbf{x}_{i^*}, \mu_{i^*}, X_{i^*}, \sum_{i=1}^{k} \pi_i^k;$$

where i^* denotes the index of the data cloud/cluster prototype/focal point. The *typicality* and *eccentricity* can be updated for the current, k^{th} data point only plus for the data cloud/cluster prototypes, centres, not for all past, k data (see steps 3 and 5 of the procedure). An important aspect is the dynamic nature of the data streams and their order dependency. One can chose to have a forgetting factor or mechanism to introduce the importance of the time instances when a particular data sample was read. This is important for data streams.

6. Classification based on *typicality*

Classification is another central element of pattern recognition, machine learning, data mining [1]. Within the TEDA framework classification can be done using *local* (instead of *global*) values for ξ, τ, ζ, and t. The main difference between the *global* and *local* expressions is the summation limits – the data samples over which the summation is performed. In the *global* case, it is performed over *all* available (by this moment in time) data samples, k. In the *local* case, the summation is over a group of data samples from a particular class or data cloud/cluster (in general, there may be more than one data cloud/cluster per class [3]); say, if we have data for healthy and ill patients, good and bad examples of something etc. we can accumulate data samples/observations for each one separately and get in this way, ξ_{good}, τ_{good}, and ξ_{bad}, τ_{bad}. Then the classifiers of zero, first or higher order can be built similarly to the AutoClass concept described in [3, chapter 8]. Zero order classifier means using the label of the classifier (singleton) as an output. First order means using a regression style classifier where for each data cloud/cluster a separate linear regression function over the input features is generated. Higher order classifiers may have non-linear output. In all cases the input part of the classifier can be seen as a clustering or data clouds (or simply data partitioning) which was described in the previous section. Zero order classifiers are more attractive from the point of view that they are easier to interpret and can be fully unsupervised [9]. First order classifiers, on the other hand, can lead to a better performance [10].

7. Prediction (and Control) Based on *Typicality*

Predictive models and controllers can be built using the multi-model principle where each *local* submodel is quite simple (e.g. linear or even zero order singleton) [3]. The problem then translates to the decomposition of the data space into (possibly overlapping) *local* regions in which often overlapping regimes and *local* behaviours is easier to define and tune. This problem on its own has been demonstrated to be possible to successfully address using clustering (or forming data clouds) in [3]. Clustering was described earlier in section *5*, therefore, due to the lack of space in this paper we will limit to just pointing towards the applicability of TEDA framework to develop and design new predictors and controllers which are not based on the traditional probability theory and, thus,

do not suffer from the limitations and assumptions on which it is based, e.g. normal or known distribution of the variables, (in)dependence of the data samples/observations, their limited (not infinite) amount, etc. Moreover, the proposed TEDA framework makes possible to extract recursively in a closed analytical form the exact distributions from the real data.

8. Conclusions

In this paper, a new systematic framework for data analytics, based on the *typicality*- and *eccentricity* of the data is proposed which is spatially-aware, non-frequentist and non-parametric. The proposed new *typicality*- and *eccentricity*-based data analytics (TEDA) framework is free from prior assumptions (such as Gaussian or any other specific distribution of the data, the need to have subjectively defined membership functions, kernels, specific proximity/distance measures, availability of infinite amount of data, independence of the data samples, etc.). Both, *typicality* and *eccentricity* can be calculated by computationally efficient recursive formulas. *Typicality* resembles fuzzy membership functions (having maximum 1) but is objectively derived from the data pattern (not due to *prior* assumptions). Normalised *typicality* resembles pdf (having a sum/integral equal to 1) but is spatially-aware. TEDA does not require any *prior* assumptions and in a more natural manner represents the real (not *purely* random) processes, such as climate, economics, industrial processes. TEDA requires no *prior* assumptions and provides close analytical expression and extracts multimodal distributions entirely from the data. It can be used for development of a range of methods for anomalies and fault detection, image processing, clustering, classification, prediction, control, filtering, regression, etc. In this paper due to the space limitations, only few illustrative examples are provided aiming proof of concept.

AUTHOR

Plamen Angelov –Intelligent Systems Research Lab, School of Computing and Communications, Lancaster University, LA1 4WA, UK.

E-mail: p.angelov@lancaster.ac.uk

REFERENCES

1. C. Bishop, *Machine Learning and Pattern Classification*, Springer, 2009.
2. D. Osherson, E. E. Smith, "Discussion: On typicality and vagueness", *Cognition*, vol. 64, 1997, pp. 189–206.
3. P. Angelov, *Autonomous Learning Systems: From Data Streams to Knowledge in Real time*, John Willey, Dec. 2012, ISBN: 978-1-1199-5152-0. DOI: http://dx.doi.org/10.1002/9781118481769
4. P. Angelov, *Anomalous System State Identification*, GB1208542.9 patent application, priority date,15 May 2012.
5. J. Iglesias et al., "Creating evolving user behavior profiles automatically", *IEEE Trans. on Knowledge Data Engineering,* vol. 24, no. 5, May 2012, pp. 854–867. DOI: http://dx.doi.org/10.1109/TKDE.2011.17
6. D. Kolev et al., "ARFA: Automated Real-time Flight Data Analysis using Evolving Clustering, Classifiers and Recursive Density Estimation". In: *Proc. IEEE Symposium Series on Computational Intelligence*, SSCI'2013, 16–19 April 2013, Singapore, ISBN 978-1-4673-5855-2/13, pp. 91–97. DOI: http://dx.doi.org/10.1109/EAIS.2013.6604110
7. L. A. Zadeh, "Fuzzy sets, *Information and Control,* vol. 8, no. 3, 1965, pp. 338–353. DOI: http://dx.doi.org/10.1016/S0019-9958(65)90241-X
8. P. Angelov**,** R.Yager, "A New Type of Simplified Fuzzy Rule-based Systems, International Journal of General Systems", vol. 41, no. 2, pp. 163–185, Jan. 2012. DOI: http://dx.doi.org/10.1080/03081079.2011.634807
9. B. S. J. Costa, P. P. Angelov, L. A. Guedes, "Fully Unsupervised Fault Detection and Identification Based on Recursive Density Estimation and Self-evolving Cloud-based Classifier", *Neurocomputing*, 2014, to appear.
10. P. Angelov, et al., "Symbol Recognition with a new Autonomously Evolving Classifier AutoClass", *2014 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS-2014*, 2–4 June, 2014, Linz, Austria, to appear.
11. http://www.martynhicks.co.uk/weather/data.php?page=m01y2014, accessed 6 February 2014
12. P. Angelov, "Fuzzily Connected Multi-Model Systems Evolving Autonomously from Data Streams", *IEEE Transactions on Systems, Man, and Cybernetics – part B, Cybernetics,* vol. 41, no. 4, August 2011, pp. 898–910.