Detecting Anomalous Behaviour using Heterogeneous Data

Azliza Mohd Ali, Plamen Angelov* and Xiaowei Gu

Data Science Group
School of Computing and Communication
Lancaster University, Lancaster, LA1 4WA, UK

Abstract. In this paper, we propose a method to detect anomalous behaviour using heterogenous data. This method detects anomalies based on the recently introduced approach known as Recursive Density Estimation (RDE) and the so called eccentricity. This method does not require *prior* assumptions to be made on the type of the data distribution. A simplified form of the well-known Chebyshev condition (inequality) is used for the standardised eccentricity and it applies to any type of distribution. This method is applied to three datasets which include credit card, loyalty card and GPS data. Experimental results show that the proposed method may simplify the complex real cases of forensic investigation which require processing huge amount of heterogeneous data to find anomalies. The proposed method can simplify the tedious job of processing the data and assist the human expert in making important decisions. In our future research, more data will be applied such as natural language (e.g. email, Twitter, SMS) and images.

Keywords: Heterogeneous data, anomaly detection, RDE and eccentricity

1 Introduction

Digital data is generated every day in exponentially growing quantity and it's become really "big data". More than 2.5 Exabyte data is being created every day and the number is doubled every few years [1]. In [2], the authors forecast that digital data will reach 16 zettabytes in 2017. Data can be seen as a raw material that can be in various forms such as text, numbers, images, video or signals. This diversity of the data leads to heterogeneous data which can also be structured and unstructured. Structured data is the data in the same format and easy to organize, while unstructured data does not have a common structure; for example, emails, images, etc. It is hard to combine and computationally analyse such type of data. In the age of big data, one of the challenges in the data analysis is how to process and integrate heterogeneous, unstructured data such as social media data, images and streaming data [3]. Extract knowledge from text and images, the social media such as email, Twitter, Facebook has become an issue because the data is in different modalities and is sometimes too short or noisy (e.g. text messaging; and WhatsApp) data [4]–[7].

The data can be seen as a raw material (e.g. facts, numbers, letters and symbols) that can be extracted from observations, experiments, computation and record keeping [8]. Until 1970 – 1980s, most of the data were scattered. Text data can be found in the documents such as letters, reports, books or journals, while image data (for example, photography was analogue, same for the radio and telephony; images were produced from negative film or drawing and signal data can be recorded using vinyl records or compact cassette and transmitted through analogue communications). All of this data had to be processed manually before it becomes useful information. The process was really hard and time consuming.

The first modern computer was introduced by John Vincent Atanasoff [9]. After this invention, many data storage technologies and the size of storage become huge [10]. For example, in 1956 a computer hard disk size was only 5 megabytes however in 2016 the size may be 10 terabytes. Following the Internet revolution in early 1990s, more data has been created from email, file sharing (FTP), and telephony (voice, fax, SMS, voice messaging). Starting in the mid-2000s, when iPhone and Android were introduced, there are a lot of applications that have been developed which create more and more data every day, especially social media applications. Now, there are many digital devices in homes, workplaces and public places, mobile, distributed and cloud computing, social media and the Internet of Things. These platforms are important to all aspects of the everyday life such as work, communication, travel and leisure and all of these generate data signature [11]. With digitalization, traditional database move to network data infrastructures and more data being publicly available make the data revolution which brought to live

*corresponding author
E-mail: p.angelov@lancaster.ac.uk

the term “big data”. In business, for example, big data is providing new resources for company activities and can leverage additional profit by enhancing productivity, competitiveness and market knowledge.

Take as an example, when crime happened, forensic investigator has to collect evidence from the crime scene and every detail about a suspect such as demography, financial or travel information to analyse it for anomalies. All possibilities have to be investigated. Before the investigator can find the real suspect, there are a lot of data and information to be processed and analysed. Usually, the investigator checks a suspect’s bank account, aiming to find abnormal transactions. The process entails checking all the bank account information such as date, transaction; debit or credit, and location. Sometimes, there are many transactions in the account and some people have two or more bank accounts. The investigator’s job becomes more tedious and time consuming if they have to check the transactions one by one. Hence, if there is a system which can process all the transactions and find the anomaly, the job becomes easier and more efficient. The system can create patterns, clusters or classes which will represent the behaviour of the suspect based on the amount of money spent. Then, the investigator has other evidence to process and find anomalies. Later, all the evidence (anomalous data) has to be integrated / fused with other data sets (evidence), creating a sequence of events. Finally the investigator can make a decision and solve the case.

Abnormal data can be detected using anomaly detection techniques [12]. Anomaly detection is one of the methods for data analytics which aims to identify the data samples that “stand out”, are “untypical”, differ from normality significantly. It can also differentiate between normal and abnormal behaviour. There are many types of problems related to anomaly detection. These include the nature of the input data, types of anomalies (points, contextual or collective anomalies), data labels (supervised, semi supervised or unsupervised) and outputs of the anomaly detection [13]. Anomaly detection is very important in analysis of fraud detection, drift detection in data streams [14] , clustering, outliers detection and autonomous video analytics, and so on. [12]. The result of such detection are used in many applications such as intrusion detection in cyber security [15], fraud detection [16], surveillance system [17] and military surveillance of enemy activities [18].

In this paper, we propose a new method to detect abnormal human behaviour using different available datasets. Datasets have been acquired from the VAST Challenge 2014 [19]. We use the feature extraction process as explained in section 4.2. The recursive density estimation (RDE) was applied and the eccentricity of each of the data samples was calculated to detect the abnormal behaviour. No *prior* assumptions of data distributions are being made. Instead, the Chebyshev inequality is being used in regards to the eccentricities of the data. This is further detailed and discussed in section 3. The rest of the paper is organized as follows. Section 2 presents the newly proposed method. Section 3 describes the anomaly detection. Then, section 4 discusses the application of the new method to the heterogeneous data. Finally, the last section concludes the paper and describes the directions of the further work.

2 Proposed method

Data may vary in terms of form (qualitative or quantitative), structure (structured, unstructured or semi-structured), producer (primary, secondary or tertiary) and type (indexical, attribute, metadata) [11]. Most data nowadays is, in reality, heterogeneous. Therefore, a combination of heterogeneous data can generate rich information insights. However, different kind of data have to be processed differently and when it became “big data”, it can be cumbersome, tedious and time consuming to process. One way to simplify the data processing part is selecting the important features in the data set through a practice known as feature extraction. It is very important pre-processing stage. When processing data, there are often outliers/anomalies. It is very important to detect and remove these first. According to [20], outlier/anomaly is defined as data points that are distant from the other agglomerated data points in the same class while [21] defines outliers as observations, which appears inconsistently in the set of data. However, outliers/anomalies can represent very valuable information, e.g. in forensic cases, as it will be demonstrated later.

Traditionally, anomaly detection is addressed using statistical methods where frequentistic technique to represent the probabilities are applied and *prior* assumption has to be made [12]. The main decision is traditionally made using a threshold values. These thresholds are based on normal distribution of random variables (usually assuming Gaussians) while for arbitrary distributions, they are based on the well-known Chebyshev inequality [12] . These approaches have the following disadvantages [12]:

a) they require strict *prior* assumptions;

b) they relax the conditions too much to avoid false positives to the level where it misses many true positives (the 3σ rule sometimes fails to detect some obvious outliers);
c) large amount of data samples is required;
d) a single data sample is compared with the average, instead of comparing pairs of data samples; therefore, the information is blurred and is no longer point-wise and local.

According to [12], eccentricity can be applied to avoid the disadvantages of the traditional statistical method. This approach does not requires any prior assumption and a σgap can be formulated between the eccentricities of the data samples with the larger eccentricity [12].

Heterogeneous data may have anomalies in each data type. All anomalies can also be combined and can create a more informative overall result e.g. per person. Data fusion can be used to enhance the decision making because it combines data from many sources. It has been widely used in multisensor environments [22]. The goal of the data fusion is to combine and aggregate the data which are derived from several sensors and these techniques can be applied to the text processing domain as well [23]. According to [23] there are three nonexclusive categories of data fusion which are; i) data association, ii) state estimation and iii) decision fusion. Data fusion is a challenging task. According to [24], there are three challenges in data fusion:
a) data is produced from very complex systems such as biological, environmental, sociological and psychological systems;
b) increased diversity, the number, the type, and scope of the data;
c) working with heterogeneous data sets means that the respective advantages of each data set are maximally exploited and drawbacks suppressed.

In this paper, we propose the idea of the automated processing of big digital data sets and streams to facilitate detection of anomalous behaviour as shown in Figure 1. This research offers a hierarchical structure of the processing data in the form of financial data (credit card and loyalty card data), signals (GPS data), natural language processing data (email, Twitter, SMS data) and image data. In this paper, only two types of data will be demonstrated without compromising the generality of the overall approach which are financial (credit card and loyalty card) and GPS data. As a first step, feature extraction will be applied followed by anomaly detection phase. All data are processed in an unsupervised manner. For example, anomaly detection based on RDE and eccentricities are applied. Next, data fusion can be applied. Because of time constraints, the data fusion will be applied in the future research. Different data modalities, such as text, images and signal data have to be integrated to form contextually linked event sequences and story lines. Then, the integrated data can be analysed, making this analysis more efficient because it will be over a significantly smaller amount of much more organised and human-intelligible information in the form of rules, graphs, clusters, and so on. The proposed approach can also formalize the existing expert knowledge and construct the sequence of events. The final step requires a human expert to verify the analysis and make the final decision based on the much smaller amount of highly intelligible information. The significance of the proposed new approach is to assist the human expert and reduce the time and to get the right conclusion at the right moment in time while, at the same time, using a much larger amount of heterogeneous data.

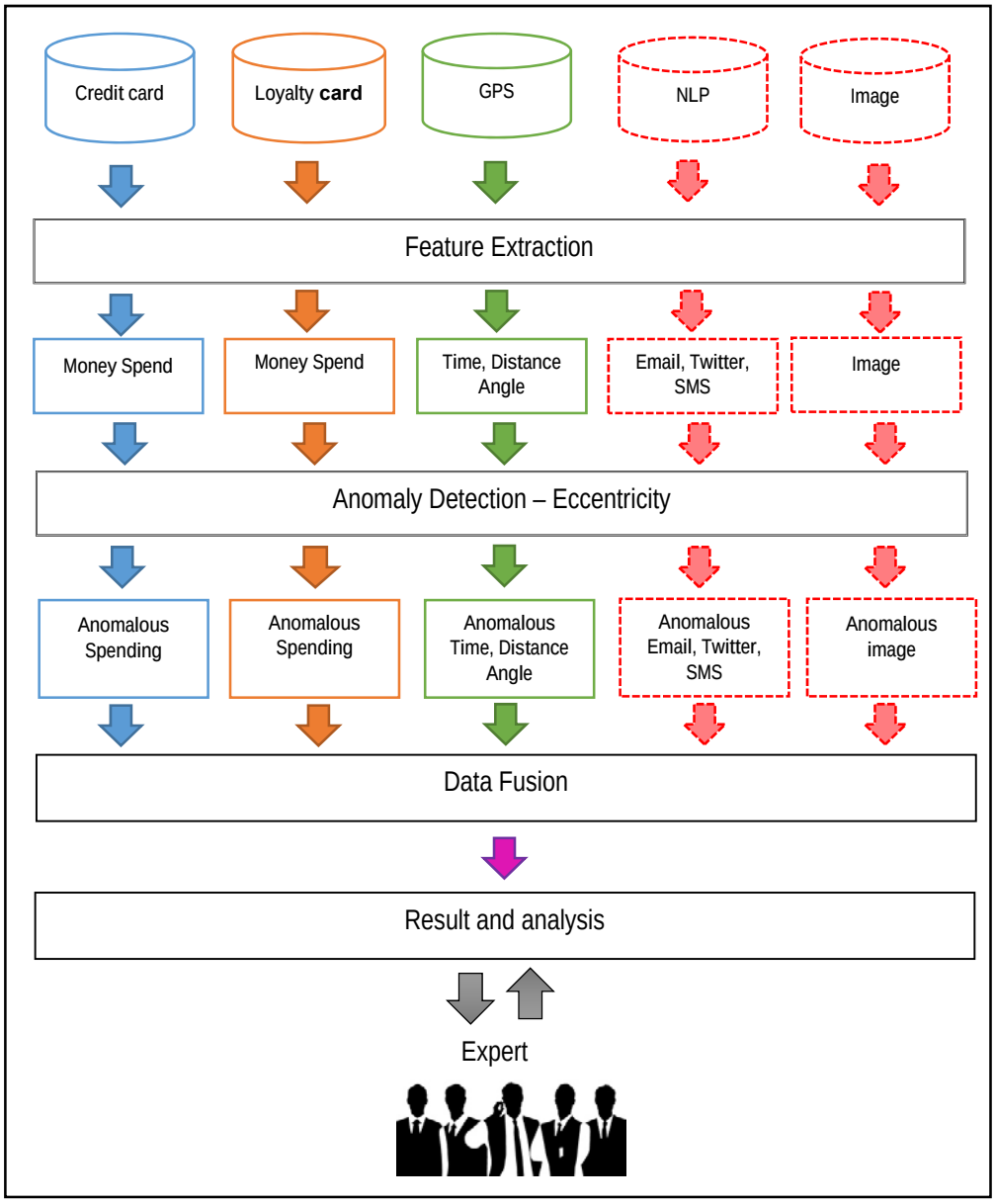


Fig. 1. The proposed method for autonomous analysis of heterogeneous data aiming to detect anomalous behaviour

3 Anomaly Detection

After the feature extraction phase, only the important features are used in the next phase which is anomaly detection. In this case, the selected features are denoted by $\{x_i\}$, where x_i denotes the financial data and $\{y_i\}$ where y_i denotes the gps data. Let us consider the data points as $\{x_i\},\{y_i\} \rightarrow \{z_i\}$. RDE is then applied to find the anomalies in the datasets. It is designed based on a Cauchy type function [25]. The most research methods assume Gaussian distribution of the data. However, obviously, real data do not necessarily follow Gaussian distribution. In RDE, the data points can be denoted as $\{z_1, z_2, ...z_k\}$, where the index k may have the physical meaning of time instant when the data item arrives. The density is calculated as follows:

$$D_{k,j} = \frac{\sum_{i=1}^{k} \pi_{k,i}}{2k\pi_{k,j}} \tag{1}$$

Accumulative proximity, π from a particular, j^{th}, $(j \geq 1)$ data point, to all remaining, $(k > 1)$ data points:

$$\pi_{k,j} = \pi_k(z_j) = \sum_{i=1}^{k} d_{i,j} \quad ; k > 1 \tag{2}$$

where $d_{i,j}$ denotes a distance between data points z_i and z_j.

A deeper analysis for detecting anomalies is using the standardised eccentricity, ε of the data samples [12]. This technique can be applied to image processing, video analytics [26], fault detection [27] and also user behaviour [28]. For instance, any data sample that has high value of the standard eccentricity $(\varepsilon_{k,j} > n^2 + 1)$ is a suspected anomaly. Where n denotes the number of sigma. The eccentricity offers a new angle of view towards the problem in comparison with the traditionally used probability [12]:

a) it is based on the data samples and their local properties in the region;
b) there are no *prior* assumptions about the distributions;
c) there is no need of a kernel;
d) there is no pre-specified user-defined, threshold or a parameter;
e) there is no need for independent, identically distributed (iid) data samples; in contrast, the eccentricity is based on their mutual dependence;
f) there is no need for unlimited number of observations and can processed with very little data sample e.g.: 3 data samples.

Because RDE framework is entirely based on the data samples and does not require any *prior* assumptions as well as problem-and user-specific parameters, we use this method for anomaly detection. The standardised eccentricity is introduced in [29] but was redefined later in [30] as the inverse of the density:

$$\varepsilon_{k,j} = \frac{1}{D_{k,j}} \tag{3}$$

The eccentricity is very useful in anomaly detection because it allows per data sample and local analysis, automatically. The eccentricity can be extracted from the data in a closed analytical form, and updated recursively. It is very useful for the research on real processes (e.g. climate, earthquakes, nuclear, tsunami and other disasters) which are often complex and uncertain and not purely random, does show inter-sample dependence, not necessarily normal/Gaussian distributions and definitely not an infinite number of observations [29]. The traditional probability theory does not work on small amount of data and for such real problems, the amounts of data are usually limited and the distributions are not normal. The so called TEDA framework [29] offers a convenient approach to easily detect anomalies and estimate the degree of severity (how bigger ε is). Eccentricity of the j^{th} data item calculated when $k > 2$ non-identical data items are available is given by:

$$\varepsilon_{k,j} = \frac{2k\pi_{k,j}}{\sum_{i=1}^{k} \pi_{k,i}} \quad ; \sum_{i=1}^{k} \pi_{k,i} > 0 \;\; ; \;\; k > 1 \tag{4}$$

where $\pi_{k,j}$ denotes accumulated proximity, π from a particular, j^{th}, $(j \geq 1)$ data point. These quantities (π and ε) 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 [29]. If use Euclidean distance [12],

$$\pi_{k,j} = k(\| z_j - \mu_k \|^2 + Z_k - \| \mu_k \|^2) \tag{5}$$

$$\sum_{i=1}^{k} \pi_{k,i} = 2k^2 \left(Z_k - \| \mu_k \|^2 \right) \tag{6}$$

where μ_k is the recursively updated (local or global) mean; Z_k is the recursively updated squared norm sum and the recursive update is made as follows [29]:

$$\mu_k = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}z_k \quad ; \mu_1 = z_1 \tag{7}$$

$$Z_k = \frac{k-1}{k}Z_{k-1} + \frac{1}{k}\| z_k \|^2 \; ; Z_1 = \| z_1 \|^2 \tag{8}$$

The standardised eccentricity can be determined by:

$$\varepsilon_{k,j} = 1 + \frac{\| z_j - \mu_k \|^2}{Z_k - \| \mu_k \|^2} \tag{9}$$

Further, in TEDA a condition which provides exactly the same result for the Chebyshev inequality without making any assumptions about the amount of data and their independence was introduced for Euclidean distance by [12], [29]:

$$P\left(\varepsilon_{k,j} > n^2 + 1\right) \leq 1 - \frac{1}{n^2} \tag{10}$$

After finding the $\varepsilon_{k,j}$, the $n\sigma$ gap principle is used to compare each of the data samples with the aim to identify the anomalies [12]:

IF ($\varepsilon_{k,j} > n^2 + 1$) THEN (z_j is an outlier) (11)

The significance of the proposed method is to assist human experts to reduce the time spent and to get the right conclusion at the right moment in time while, at the same time, allowing access to a huge amount of data. Such an approach can shorten the pre-processing phase and increase the efficiency of the use of a human expert.

3.1 Data Fusion

In this paper, the data fusion will not be applied. In our future research the data fusion will be applied to produce better result and analysis.

4 Applying the New Method to the Heterogeneous Data from the VAST 2014 Challenge

In this paper, we consider this popular example as an illustration only and as a proof of concept without limiting the overall methodology.

4.1 Datasets

The data was acquired from the IEEE Visual Analytics Science and Technology (VAST) Challenge 2014 [19]. In this challenge, there are four datasets taken from 6 January to 19 January 2014. The description of the data is defined in Table 1. The GPS data is transformed into direction, average speed, distance and ratio of trajectory angle. Truck drivers' data is removed because there are missing values for the IDs of truck drivers.

Table 1. Description on Datasets

Datasets	No. of data points	Attributes
1. Credit Card	1492	1. Timestamp 2. Location 3. Price 4. First Name 5. Last Name
2. Loyalty Card	1393	1. Timestamp 2. Location 3. Price

		4. First Name 5. Last Name
3. GPS Data	685170	1. Timestamp 2. Car ID 3. Latitude 4. Longitude
4. Car Assignment	45	1. First Name 2. Last Name 3. Car ID 4. Current Employment Type 5. Current Employment Title

4.2 Feature Extraction

Without being limited to this specific data set, we consider the types of data that are available in the VAST 2014 challenge.

a) Credit card and loyalty card data

This includes financial data concerning the money spent by staff members. Transactions of credit cards will also appear in the loyalty card if they are swept together. Normally, all transactions are the same except when they are not using the loyalty card or maybe the shop did not accepted the card. Therefore, for some of the transactions the two values are not the same. In the VAST 2014 Challenge data that we consider, there are five attributes (timestamp, location, money spend, first name and last name) in credit card and loyalty card data but only money spent is extracted from these datasets. Money spent according to the credit card, C_i and money spent according to the loyalty card, L_i .

$$x_i = [C_i, L_i] \tag{12}$$

where i denotes the data points. From these datasets, different features can be extracted including:

1. Total spending per person.
2. Total spending per person and per day.
3. Total spending per location.
4. Total spending per location per day.

These features can be extracted from the credit card and loyalty card data. Features including “total spending per person” can be extracted by totaling up every spending on credit card and loyalty card for every staff member. Therefore, we can have 45 data points because there are 45 staff members. From this data, it will be easy to see which staff member spends more and which staff member spends less. Then, total spending by every staff member per day can be extracted. This can show which day the staff member spends more or less. After that, the total amount of spending per location can be extracted. From the location, we can find in which place people spend more. The last feature is creating data for every location per day. Hence, it will be easier to find which day and which location is significantly different from others. These features can give results on which person has a suspicious spending behaviour, when the suspicious behaviour of spending happens and where the suspicious spending took place. Money spent is normalized between 0 and 1 to make the data comparable. Normalization requires the range (min, max) per feature:

$$x_{norm} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{13}$$

b) GPS data

The GPS data has four attributes which are;

i) Timestamps,
ii) ID,
iii) GPS coordinates (longitude and latitude).

The most important features are the GPS coordinates. Trajectory information can be determined from the GPS coordinates such as projection of the trajectory, average speed, ratio of the trajectory angle and distance. Trajectory can represent mobility of people (e.g. people moving by bicycle or jogging carrying mobile phone), mobility of transportation (e.g. vehicle supported by GPS – taxis, bus, aircraft) mobility of animals (e.g. biologist collecting the moving trajectories animal – migrating or behaviour) and mobility of natural phenomena (e.g. meteorologists, environmentalists, climatologists – collecting trajectories of some natural phenomena – hurricanes, tornados) [32]. After the pre-processing, the GPS data can be compressed as:

$$y_i = [\, N_i, d_i, \bar{R}_i \,] \tag{14}$$

In [33] several steps were proposed to get different features from the trajectory type data.

i. Trajectory, T_i

$$T_i = \left\{ \left(a_{i,j}, b_{i,j} \right);\ j = 1, ..., N_i \right\} \tag{15}$$

Where $a_{i,j}$ denotes the latitude and $b_{i,j}$ denotes the longitude, N_i is duration (in seconds) of the trajectories contain in that specific sample.

ii. Projections of the trajectory, r_i

Projections include the horizontal and vertical projections of the trajectory. This shows the trajectory of a person from start point to end point of the destination. The advantage of using this feature is to separate the trajectory per axis. Projections r_i are defined as:

$$r_i = \left(a_{i,N_i} - a_{i,1}, b_{i,N_i} - b_{i,1} \right) \tag{16}$$

iii. Average Speed, v_i

The next feature is the average speed. It is derived from the distance of the route. This feature can also differentiate vehicles with varying speed. Average speed, v_i is calculated as:

$$v_i = \frac{1}{N_i - 1} \sum_{i=1}^{N_i - 1} (a_{i,j+1} - a_{i,j}, b_{i,j+1} - b_{i,j}) \tag{17}$$

iv. Distance, d_i

The distance shows the span of the destination from the start point till the end point is. The distance, d_i is calculated as:

$$d_i = N_i v_i \tag{18}$$

v. Ratio of the trajectory change angle, $\bar{R}_i$

The trajectory angle $\theta_{i,j}$ is calculated to find the sharpness of turns in the trajectory, T_i. The angle $\theta_{i,j}$ is then compared to check whether it is less than 90^o or not. 90^o is chosen to compare the angle because normally when people move, they will go straight to destination and sometimes turn back but not always. The trajectory may reach

90^o but it is quite abnormal if we turn more than 90^o. R_i denotes the time the angle exceeds 90^o during single trajectory. If the angle is less than 90^o, then it will give the number of normal values, $R_i \leftarrow R_i + 0$, else the number of abnormal values, $R_i \leftarrow R_i + 1$. Then, the ratio $\bar{R}_i$ is calculated. The trajectory change angle, $\theta_{i,j}$ and the ratio $\bar{R}_i$ are defined as follows:

$$\theta_{i,j} = \arctan\left(b_{i,j+1} - b_{i,1}, a_{i,j+1} - a_{i,1}\right) \tag{19}$$

$$\text{IF } (\theta_{i,j} < 90^\circ) \text{ THEN } (R_i \leftarrow R_i + 0) \text{ ELSE } (R_i \leftarrow R_i + 1) \tag{20}$$

$$\bar{R}_i = \frac{R_i}{N_i} \tag{21}$$

All the features are normalized to make the data comparable. The process is similar to the equation (13) used for the credit and loyalty card data.

c) Natural language processing (NLP) and image data

NLP is a computer science fields which analyses how humans interact with the computer. It helps interpret human language and translate it to the computer and digital data. Recently, much research on social media involving NLP was made. Social media was introduced in 2003. Now, there is more social media data available such as email, Twitter, Facebook and WhatsApp. These data is unstructured and requires the NLP technique to pre-process the data. Features such as keywords, topics, etc. need to be extracted from the data.

Image data differ from text data. Digital images consist of binary representations. Many formats are available such as jpg, bmp, gif and png. The size of the digital images is based on the number of pixels. There are several steps to be done in pre-processing digital images such as image resampling, segmentation, grey scale and noise removal. In this paper, we will not cover these two data types mainly because of the lack of space and time. These data type will be used, however, in our future research.

4.3 Case Study

Figure 2 shows the standardised eccentricity, ε_{jk} for the money spent based on the credit and loyalty card data. There is one noticeable anomaly in this data where the staff member no. 31 spends 10,000. Other attributes have been also analysed (money spent, day and staff member). This figure clearly shows that only one staff member spent too much, which is abnormal. Staff member no.31 spent an obviously high amount in one day compared to the other 13 days and other staff members. Then, we remove the abnormal data which is 10,000 (see Figure 3). The result shows 2 anomalies in this data. There is one case which shows an obvious anomaly (spending of 600 by staff member no. 43) which is $> 5\sigma$ away from the mean. According to the Chebyshev inequality, this translates to <4% of the data. Figure 4 shows that there is one anomaly in the loyalty card data. The same staff member no. 43 spends 600 and is the noticeable one. Again, this is $> 3\sigma$ while the other anomalies are above 5σ. When we analyse daily spending on credit card and loyalty card data, it shows two obvious spending patterns by staff member no. 43 and staff member no. 40. The two staff members are the top management of this company where staff member no. 43 is the CEO (Chief Executive Officer) of the company and staff member no. 40 is the COO (Chief Operating Officer). Therefore, they have the power of spending a lot of money. However, the suspicious thing is that, staff member no. 31 spends 10,000 on the credit card but is not listed in the loyalty card data for this transaction. Normally, when people spend too much they will swipe their loyalty card together with the credit card. We can make an assumption that maybe (s)he is not bringing the loyalty card when spending 10,000. When analyzing the location, where the money were spend, it shows clearly in figure 5, that the highest credit card spending was at Frydos Autosupply but the loyalty card is just half of the spending of the credit card. Compared to other locations, all the spending using the credit card are almost the same with the loyalty card. This is again to show the suspicious spending using credit card for staff member no.31. Figure 6 shows the total spending of every staff member using credit card and again staff member no.31 spends obviously highest compared to the other staff members. For the loyalty card, the spending pattern is almost the same and there is no anomaly detected in this dataset (see Figure 7). For all staff

members' the spending patterns was analysed for each day. Figures 8 and 9 show the total spending per day for staff member no. 31 using credit and loyalty card, respectively. It shows that, the highest spending for this staff member is in the day 8 and there is no anomaly in the loyalty card. After that, every location has to be analysed. Location no. 11, Frydos Autosupply has the highest number of money spent in this location using credit card but there is no difference with the loyalty card. Based on this analysis and discrepancies, we can make a conclusion about the credit card and loyalty card spending behaviour. Staff member no. 31 has a suspicious behaviour based on the spending using the credit card on day 8 at the location no. 11, Frydos Autosupply. Nevertheless, the spending behaviour using loyalty card is normal and there is no suspicious behaviour detected in this dataset.

A possible explanation can be that someone else but not staff member no.31 used his/her credit card. This can reduce the huge amount of raw data into a much smaller amount of suspicious data (in this case, regarding staff members no.31, 40 and 41) and location, Frydos Autosupply which maybe further clarified if use also video from the CCTV (if available). As it will be demonstrated later, this can also be identified by analyzing the travel data.

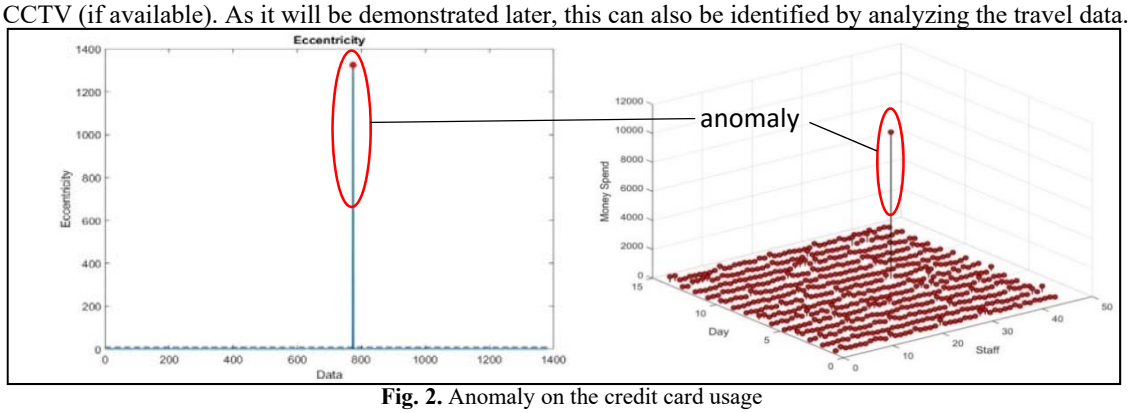


Fig. 2. Anomaly on the credit card usage

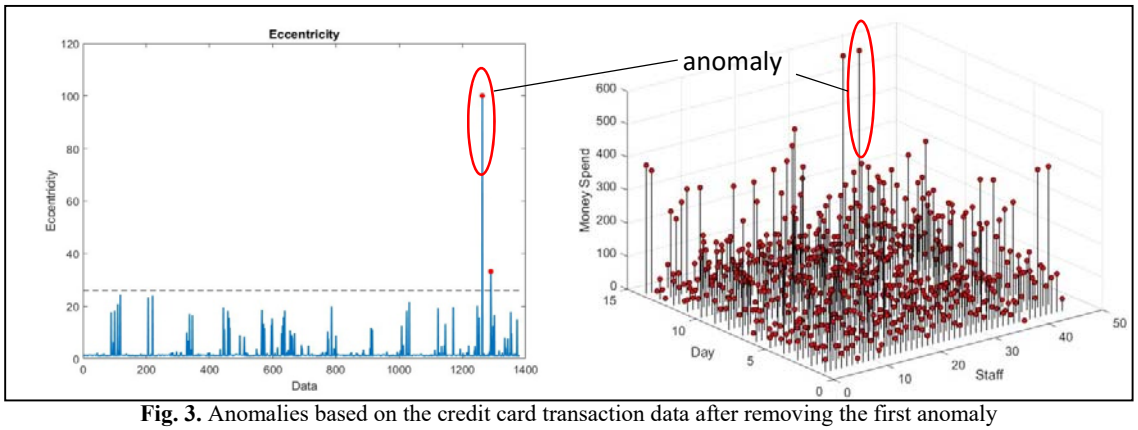


Fig. 3. Anomalies based on the credit card transaction data after removing the first anomaly

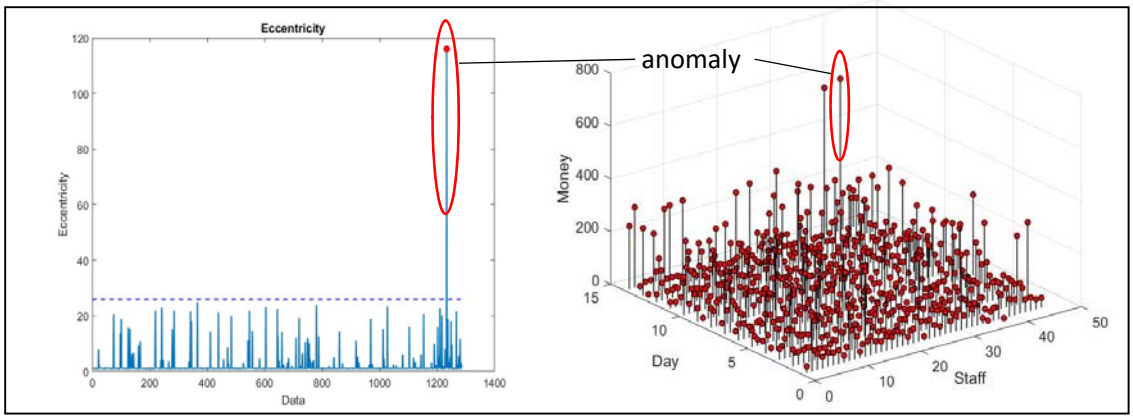
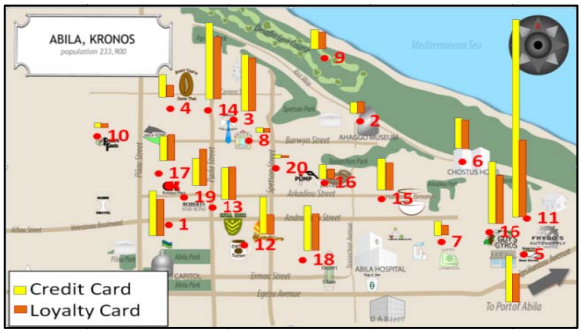


Fig. 4. Anomalies on the basis of the loyalty card data



Locations			
1	Abila Zacharo	11	Frydos Autosupply n' More
2	Ahaggo Museum	12	Gelatogalore
3	Albert's Fine Clothing	13	General Grocer
4	Bean There Done That	14	Hallowed Grounds
5	Brew've Been Served	15	Jack's Magical Beans
6	Chostus Hotel	16	Katerina's Café
7	Coffee Cameleon	17	Ouzeri Elian
8	Coffee Shack	18	Shoppers' Delight
9	Desafio Golf Course	19	Roberts and Sons
10	Frank's Fuel	20	U-Pump

Fig. 5. Comparison of the total spending using credit card and loyalty card in different locations

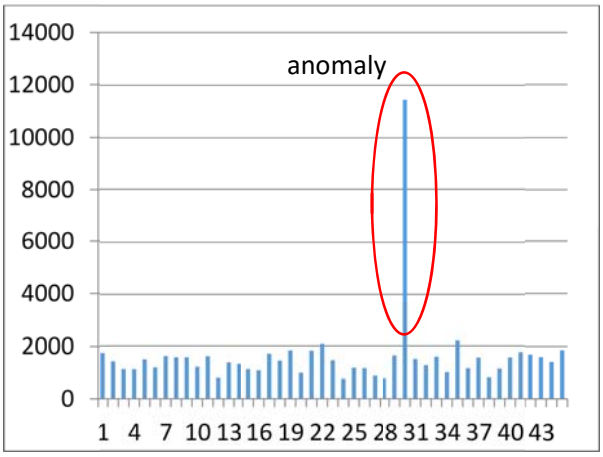


Fig. 6. Total spending per person using credit card data

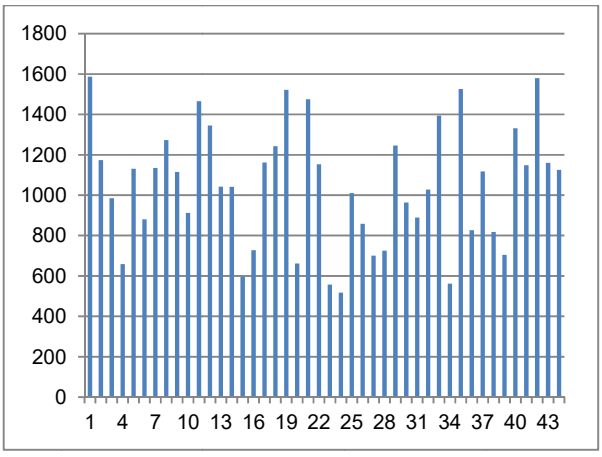


Fig. 7. Total spending per person using loyalty card data

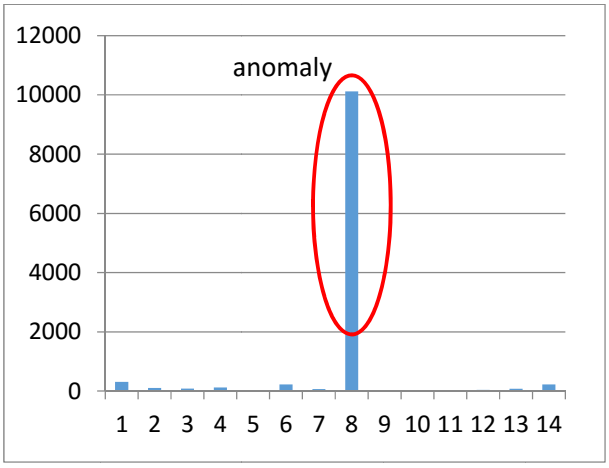


Fig. 8. Total spending per day using credit card data- Staff Member no.31

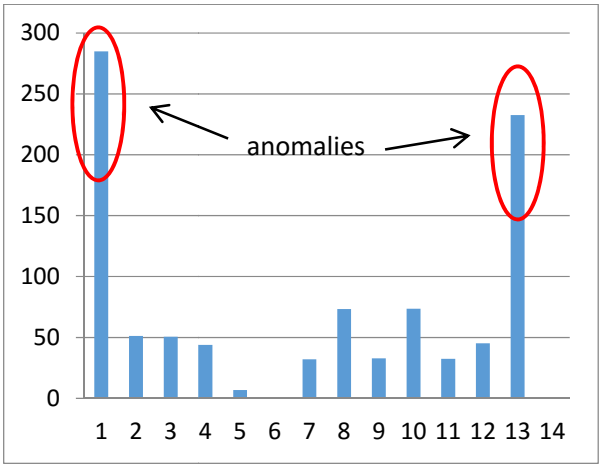


Fig. 9. Total spending per day using loyalty card data - Staff Member no.31

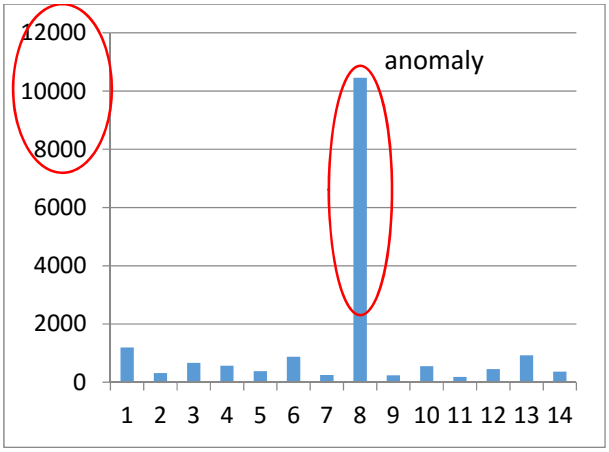


Fig.10. Total spending using credit card data per day at Frydos

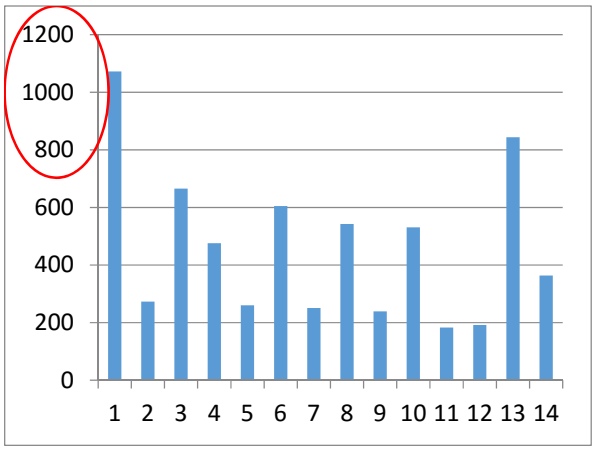


Fig. 11. Total spending using loyalty card data per day at Frydos

The data about the GPS position of the vehicles used by staff members is the biggest dataset and gives a lot of useful information about the trajectory of their movement. All the information about the travel time, travel distance and trajectory angle ratio have been transformed and used to calculate eccentricity to discover the suspicious behaviour among the staff members. Figure 12 shows the eccentricity based on the travel time, distance and ratio of the trajectory angle. The anomalous data shows that there is one staff member that travelled 14 times between 6 January and 19 January 2014. The anomalies are concerning staff member no. 18. Figure 13 also shows examples of patterns of both normal and abnormal travel behaviour. Figure 14 shows the comparison of the eccentricity of travel time, distance and ratio of the trajectory angle. Table 2 shows the ID, date and travel behaviour (abnormal). In this table, all of the travel behaviours are abnormal. Related to the suspicious spending from the credit card, the analysis of the trajectory is made on the day of the transaction. The credit card of staff member no.31 has been charged by 10,000 on 13 January 2014 at 19.20 pm. Figure 15 shows a comparison of the trajectory of the staff member no.31 and staff member no.41. The trajectory is shown from 17.57 pm to 20.10 pm on 13 January 2016. The trajectory shows that staff member no.31 did not go to the location where the credit card has been charged while staff member no.41 has a trajectory, to the location at the same time the credit card has been charged. It shows how we automatically using the newly proposed data analysis method detected something suspicious for staff member no.41. After the highest amount of spending on 13 January 2014, staff member no.31 is not using the credit card until 16 January 2014. We assume the credit card is not with him/her from 13 January 2014 to 15 January 2014. (S)he started using the credit card again only after 16 January 2014. A possible explanation can be that staff member no.31 went somewhere else and left his/her credit card which was misused by staff member no.41.This was detected fully automatically and also it was determined that this amount spent has extremely high value of eccentricity ($> 35\sigma \rightarrow < 0.1\%$ of data samples)

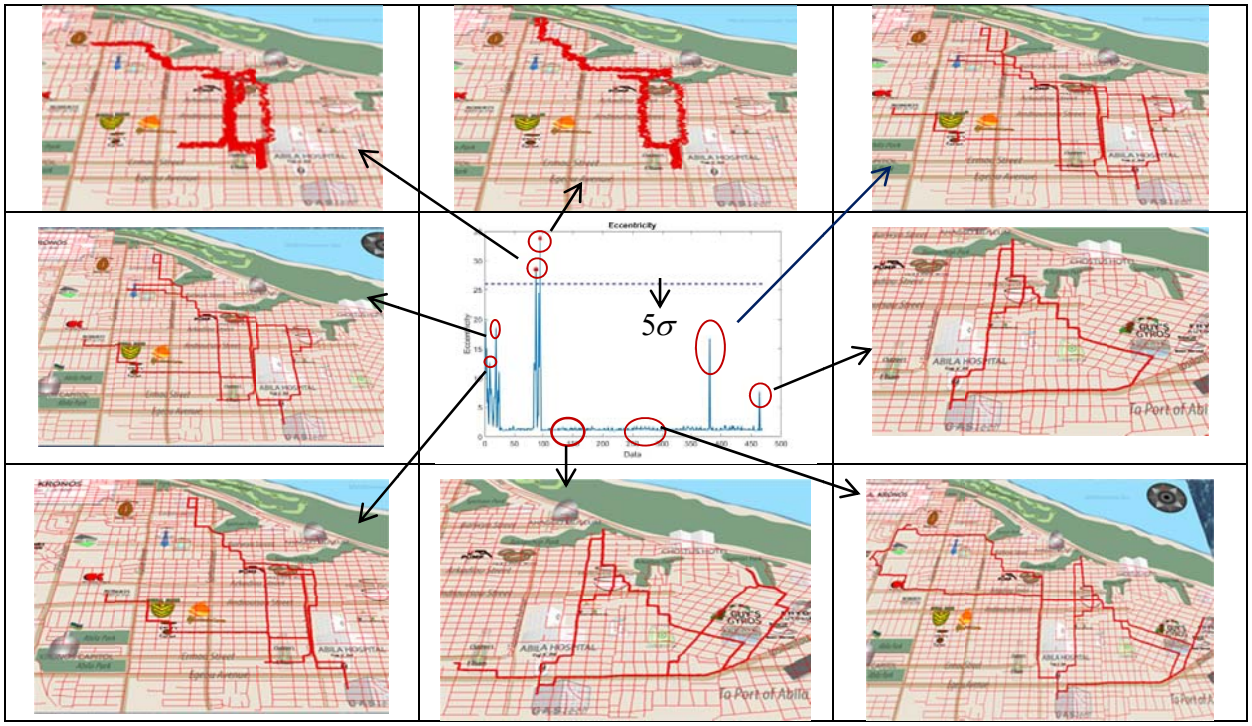


Fig. 12. Eccentricity on travel time, distance and trajectory angle ratio and normal and abnormal behaviour.

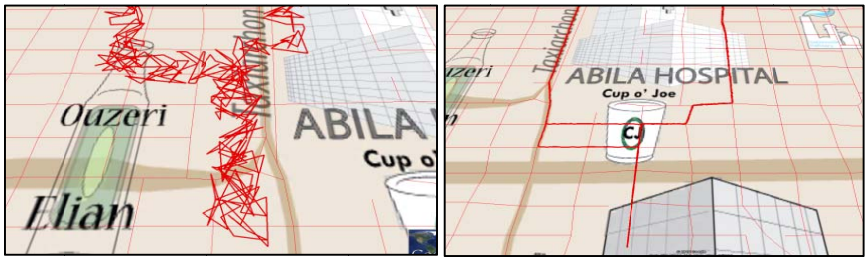


Fig.13. Example of abnormal and normal trajectory

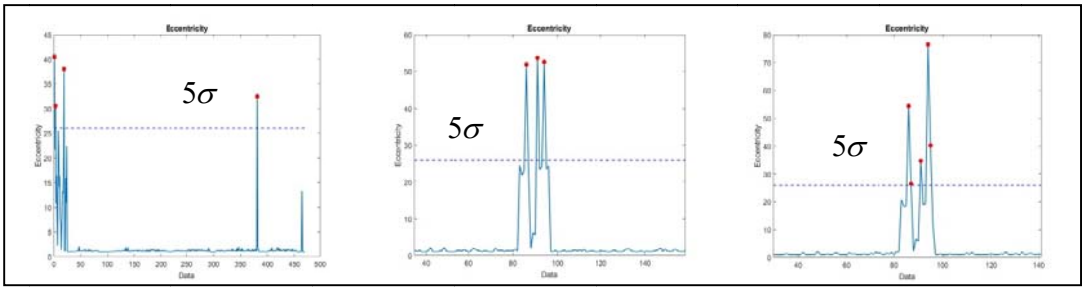


Fig. 14. Anomalies detected based on the travel time, distance and ratio of the trajectory angle

Table 2. Description on Datasets

ID	Date	Normal/Abnormal
18	9 Jan 2014	7 trips – abnormal
18	17 Jan 2014	7 trips – abnormal

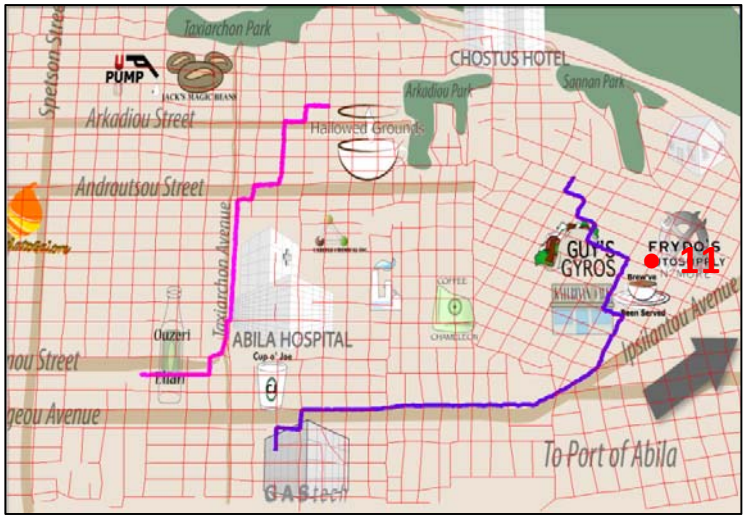


Fig. 15. Comparison of the trajectory for staff member no.31 (left) and staff member no.41 (right)

5 Conclusion

In this paper, we propose a method to detect anomalous behaviour based on heterogeneous data streams fully autonomously. This method is based on the use of the RDE and the eccentricity of each data sample. The illustrative example using VAST 2014 Challenge data [19] shows that anomalous behaviour can be detected from the dataset autonomously. There are three data sets which are credit card, loyalty card and GPS data. Anomaly is detected in every data set. There is one person consistent in all the anomalies. After analysed the anomalous data on trajectory and comparing all information from the three data sets, the new suspicious person is discovered. Therefore, this method can assist the human experts in simplifying their job and helping them in making decisions. In our future work, a variety of datasets will be used such as social network data, image or streaming data. Then, we plan to apply data fusion among the datasets and get results of more complex real problems. The results can be in the form of rules or sequence of event which later can assist the human expert and make their job more efficient. This will be subject of our further research.

Acknowledgement

The first author would like to acknowledge the support from the Ministry of Education Malaysia and Universiti Teknologi MARA, Malaysia for the study grant. The second author would like to acknowledge the New Machine Learning Methods grant from The Royal Society (Grant number IE141329/2014).

References

[1] Ernst & Young, “Forensic Data Analytics,” 2013.
[2] IDC, “Where in the World is Storage: A Look at Byte Density Across The Globe,” 2013.
[3] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, “Big data and its Technical Challenges,” *Commun. ACM*, vol. 57, no. 7, pp. 86–94, 2014.
[4] D. Turcsany, A. Bargiela, and T. Maul, “Local Receptive Field Constrained Deep Networks,” *Inf. Sci. (Ny).*, vol. 349–350, pp. 229–247, 2016.
[5] B. J. C. Principe and R. Chalasani, “Cognitive Architectures for Sensory Processing,” *Proceeding IEEE*, vol. 102, no. 4, 2014.
[6] S. Maldonado and G. L’Huillier, “SVM-Based Feature Selection and Classification for Email Filtering,” *Pattern Recognit. - Appl. Methods*, vol. 204, pp. 1–11, 2013.
[7] P. Angelov and P. Sadeghi-Tehran, “A Nested Hierarchy of Dynamically Evolving Clouds for Big Data Structuring and Searching,” *Procedia - Procedia Comput. Sci.*, vol. 53, pp. 1–8, 2015.
[8] C. L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure and the Internet*. The MIT Press, 2007.

[9] J. Vincent, “Advent of Electronic Digital Computing,” *IEEE Ann. Hist. Comput.*, vol. 6, no. 3, pp. 229–282, 1984.
[10] L. Mearian, “Data Storage: Then and Now,” *Computerworld*, 2014.
[11] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* SAGE Publications Ltd, 2014.
[12] P. Angelov, “Anomaly Detection based on Eccentricity Analysis,” *2014 IEEE Symp. Ser. Comput. Intell.*, 2014.
[13] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
[14] E. Lughofer and P. Angelov, “Handling drifts and shifts in on-line data streams with evolving fuzzy systems,” *Appl. Soft Comput. J.*, vol. 11, no. 2, pp. 2057–2068, 2011.
[15] H. Om and A. Kundu, “A hybrid system for reducing the false alarm rate of anomaly intrusion detection system,” *2012 1st Int. Conf. Recent Adv. Inf. Technol. RAIT-2012*, pp. 131–136, 2012.
[16] Y. Kim and A. Kogan, “Development of an Anomaly Detection Model for a Bank’s Transitory Account System,” *J. Inf. Syst.*, vol. 28, no. 1, pp. 145–165, 2014.
[17] B. Delgado, K. Tahboub, and E. J. Delp, “Automatic Detection of Abnormal Human Events on Train Platforms,” no. 2009, pp. 169–173, 2014.
[18] Y. Wu, A. Patterson, R. D. C. Santos, and N. L. Vijaykumar, “Topology Preserving Mapping for Maritime Anomaly Detection,” pp. 313–326, 2014.
[19] “VAST Challenge 2014,” 2014. [Online]. Available: vacommunity.org/VAST Challenge 2014.
[20] M. Kang, R. Islam, J. Kim, J. Kim, and M. Pecht, “A Hybrid Feature Selection Scheme for Reducing Diagnostic Performance Deterioration Caused by Outliers in Data-Driven Diagnostics,” vol. 63, no. 5, pp. 3299–3310, 2016.
[21] D. M. Hawkins, *Identification of Outliers*. Chapman & Hall, 1980.
[22] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion : A review of the state-of-the-art,” *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.
[23] F. Castanedo, “A Review of Data Fusion Techniques,” *Sci. World J.*, vol. 2013, 2013.
[24] D. Lahat, T. Adali, and C. Jutten, “Multimodal Data Fusion : An Overview of Methods , Challenges , and Prospects,” *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
[25] P. Angelov, “Evolving Fuzzy Systems,” *Comput. Complex. Theory, Tech. Appl.*, vol. 2, no. 2, pp. 1053–1065, 2012.
[26] P. Angelov, R. Ramezani, and X. Zhou, “Autonomous Novelty Detection and Object Tracking in Video Streams using Evolving Clustering and Takagi-Sugeno type Neuro-Fuzzy System,” pp. 1456–1463, 2008.
[27] B. S. J. Costa, P. P. Angelov, and L. A. Guedes, “Real-time fault detection using recursive density estimation,” *J. Control. Autom. Electr. Syst.*, vol. 25, no. 4, pp. 428–437, 2014.
[28] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, “Creating evolving user behavior profiles automatically,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 854–867, 2012.
[29] P. Angelov, “Typicality Distribution Function - A New Density - based Data Analytics Tool,” in *IJCNN 2015 International Joint Conference on Neural Networks*, 2015.
[30] P. Angelov, G. Xiaowei, D. Kangin, and J. Principe, “Empirical Data Analysis: A New Tool for Data Analytics,” in *IEEE International Conference on Systems, Man, and Cybernetics*, 2016.
[31] P. Angelov, “Outside the Box: An Alternative DAta Analytics Frame-work,” *J. Autom. Mob. Robot. Intell. Syst.*, vol. 8, pp. 35–42, 2013.
[32] Y. U. Zheng, “Trajectory Data Mining : An Overview,” vol. 6, no. 3, pp. 1–41, 2015.
[33] P. Sadeghi-Tehran and P. Angelov, “A Real-time Approach for Novelty Detection and Trajectories Analysis for Anomaly Recognition in Video Surveillance Systems,” in *2012 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2013, pp. 108 – 113.