Soft Computing (2018) 22:3187–3201
https://doi.org/10.1007/s00500-017-2989-5

FOCUS

Anomalous behaviour detection based on heterogeneous data and data fusion

Azliza Mohd Ali[1,2] · **Plamen Angelov**[1,3]

Published online: 6 January 2018

Abstract
In this paper, we propose a new approach to identify anomalous behaviour based on heterogeneous data and a data fusion technique. There are four types of datasets applied in this study including credit card, loyalty card, GPS, and image data. The first step of the complete framework in this proposed study is to identify the best features for every dataset. Then, the new anomaly detection technique which is recently introduced and known as empirical data analytics (EDA) is applied to detect the abnormal behaviour based on the datasets. Standardised eccentricity (a newly introduced within EDA measure offering a new simplified form of the well-known Chebyshev inequality) can be applied to any data distribution. Image data are processed using pre-trained deep learning network, and classification is done by using support vector machine. Most of the other data used in our previous work are of type "signal"/real number (e.g. credit card, loyalty card and GPS data). However, a clear conclusion that a misuse was made very often cannot be reached based on them only. When gender or age is different from the expected, it is obvious misuse. At the final stage of the proposed method is combining anomaly result and image recognition using data fusion technique. From the experiment results, this proposed technique may simplify the tedious job in the real complex cases of forensic investigation. The proposed technique is using heterogeneous data which combine all the data from the VAST Challenge as well as image data using an introduced data fusion technique. These can assist the human expert in processing huge amount of heterogeneous data to detect anomalies. In future research, text data can also be used as a part of heterogeneous data mixture, and the data fusion technique may be applied to other datasets.

Keywords Heterogeneous data · Anomaly detection · Image processing · Data fusion

1 Introduction

Data are a raw material which may be available offline and online. In the 1970s to 1980s, most data were recorded in files, reports or books, and they were scarce. During that time, data were hard to process, and this was time-consuming. Now, most of the data are digital and easy to access. We are now talking about "big data" that is generated every day and growing exponentially. According to International Data Corporation, UK (IDC 2013), data will reach 16 zettabytes in 2017. Data is created from sensors, social media such as Facebook or Twitter, smartphone applications such as GPS signal, digital pictures and video and purchase transaction records. Heterogeneous data such as text and images from social media are the example of unstructured data which have different modalities: sometimes too short and containing noise. Due to the enormous amount of data, data scientists processed all of these data to produce beneficial results which can assist experts in making decisions. Data-driven method is a new area of study which helps to process and extracting knowledge and information in various forms (Dhar 2013). For example, when people shop and buy grocery at the supermarket, the supermarket gives their customer's loyalty card. Then, customers disclose their personal information to the supermarket which, in turn, can create customer's spending behaviour profile from shopping lists by using data min-

Communicated by P. Angelov, F. Chao.

✉ Plamen Angelov
p.angelov@lancaster.ac.uk

Azliza Mohd Ali
a.mohdali@lancaster.ac.uk

[1] School of Computing and Communications, Lancaster University, Lancaster LA14WA, UK

[2] Faculty of Computing and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[3] Honorary Professor, Technical University, Sofia 1000, Bulgaria

ing techniques. Special offers are being given on selected items that are always purchased, and this makes customers happy and loyal to the supermarket. Moreover, there are many ways to detect anomalous human behaviour which can help investigators to figure out details of any abnormal situation immediately after it happens.

Human behaviour can be identified from everyday routine. Data can be created from mobile applications such as WhatsApp and Telegram. All information about communication in the applications can be processed to produce a pattern of communication. Another example is the travel pattern. Travel data can be derived from Global Positioning System (GPS) to show people's travel pattern. All of this pattern data will give information about the normal human behaviour. Detecting human behaviour is important, especially for the investigators to exemplify when a crime took place. Forensic investigators have to investigate all the possible evidence to find the suspect. Data can be collected from surveillance cameras, social media accounts, telephone calls, and credit card purchases. Surveillance camera produces video and images, social media data may produce text, images and video, a phone call produces signal data, and credit card purchases exemplifiers' financial data. In reality, most of the gathered data produce a pattern of normal human behaviour. However, if there is something suspicious the data, then we can identify the anomalous human behaviour. Different anomaly detection methods can be applied to identify the anomalous human behaviour.

Anomaly detection is a method or process to identify data samples, which differs from the normal behaviour. In statistics, anomaly or outliers are removed to get a better result of the analysis (Pollet and van der Meij 2016). Anomaly or outliers are important in some applications. Anomaly data can help solving the problem and giving a better solution for certain problems in forensic applications, medical, surveillance systems, and much more. This method has been successfully applied in:

(a) Intrusion detection systems (IDS)—IDS can monitor cyber-attacks or cyber threats in the network systems and server applications. Anomaly detection evaluates monitoring data against normal baseline and will issue an alert if there is an occurrence of the abnormal behaviour. Challenges in IDS are the big heterogeneous data which need to be processed in real-time (Zuech et al. 2015). Zuech et al. (2015) stated that correlating security events from various heterogeneous sources such as network and server could enhance the cyber threat analysis and cyber intelligence. Machine learning can be used to learn the nature of the normal traffic behaviour autonomously, which can adapt normal structure and recognise suspicious or anomalous events (Palmieri et al. 2014). This technique can detect unknown attacks and not be restricted to any specific environment.
(b) Fraud detection—Anomaly detection has been successfully applied in financial application to detect fraud (Kim and Kogan 2014). Fraud detection is important, especially in the era of electronic commerce which involves electronic payment systems (Abdallah et al. 2016). Digital transactions always carry a risk, and the scammers are hard to be identified. Behavioural profiling method can model each behavioural pattern and detect the abnormal behaviour in a transaction (Jyothsna 2011). Credit card fraud detection also applies the same technique, which is after the cardholder profile is created; the system can analyse spending behaviour patterns of the user. If an inconsistency appears throughout the transaction, then the suspicious activity can be detected (Malekian and Hashemi 2013).
(c) Medical applications—Patient monitoring such as electrocardiography (ECG) is an example of a medical application that utilises anomaly detection. ECG is a test that produces a signal from the heartbeat to monitor possible heart problems (Keogh et al. 2006). The reported application also develops a new algorithm based on time series to detect an abnormality in the ECG signal. However, the algorithm cannot be applied to large datasets in real-time interaction. Meanwhile, Salem et al. (2013) also detected an anomaly in ECG, but they created a combination with various patient metrics such as blood pressure, body temperature, respiration rate, and blood glucose level. This study develops a wireless sensor network to detect an anomaly in the patient's body and achieved high detection accuracy.
(d) Surveillance systems—Surveillance system uses closed-circuit television (CCTV) and is installed in the buildings for security purposes. The system analyses suspicious movement recorded in the video and categorised them as an abnormal activity. Examples of surveillance system applications have been reported in Delgado et al. (2014) and Li et al. (2014). A surveillance system applied to the train platform to monitor people jumping or falling off train platform is reported in Delgado et al. (2014), which can be considered as an abnormal activity, while Li et al. (2014) created a model of the crowded scene and applied benchmark dataset to detect an anomaly in the pedestrian walkway. These two applications were utilised in crowd environment and Delgado et al. (2014) claim their system to be capable of achieving 90% accuracy in detecting an anomaly. Li et al. (2014) achieved a better result as compared with other techniques.
(e) Maritime surveillance systems—Anomaly detection was applied in maritime surveillance systems to assist in finding abnormal behaviour in the trajectory of vessels. Maritime surveillance systems used Zactivities. In Wu

et al. (2014) authors apply topology-preserving mapping (TPM) and Pallotta et al. (2013) applied rule-based and log-likelihood methods to detect an anomaly. These are unsupervised methods. TPM can capture and visualise vessels' behaviour, and have probability estimator which later can evaluate the likelihood and detect the anomaly.

From the literature, most of the detected anomalies is regarding suspicious behaviour. Data on suspicious behaviour can be derived from many sources or sensors such as credit card spending, travel information from GPS signal, medical records (based on medical checkup results), and video in public places (video surveillance). These heterogeneous data can be combined to produce more comprehensive information or knowledge. For example, data from the most popular social network, Facebook, information which can be in the forms of text, image or video, and sometimes tag of location, based on which signal can be identified. Data fusion is one of the techniques to combine and integrate data and create more accurate information by enhancing decision-making. Data imperfection, data association, data correlation, and data dimensionality are challenging factors when applying data fusion (Khaleghi et al. 2013). However, the most challenging part of data fusion is working with heterogeneous data (Lahat et al. 2015). Heterogeneous data will have multiple features. Obtaining reliable methods for fusing has become an issue and problem (Bakshy et al. 2012). Therefore, in this study, we introduce a data fusion technique which can combine heterogeneous data and contribute to making the decision.

In this paper, and even more in a real situation, when the amount of data (especially video/images but also bank transactions) will grow exponentially, there is indeed a big data problem. The data are huge (we only use the small example in the paper), and in real time such as navigation, finance, social network and internet of things, processing time is important (Philip Chen and Zhang 2014). The amount of data that a human expert is confronted with can be reduced by detecting the anomalous data and allowing the human expert to focus on these anomalous data instances only. We propose a data fusion technique to combine information about anomalous behaviour from heterogeneous data. The proposed method helps to address the variety in the big data which refers to the heterogeneity of the data. The case study considered in this paper deals with an example dataset for financial transactions. The images and GPS data can be considered as big data. The GPS data represents every second of the car's movements. The proposed method can help managing real situations. For example, real-time data streams such as bank transactions, CCTV images which can be processed recursively using the algorithm in the proposed method. The proposed approach adds value by allowing the human to reach a high throughput in processing huge amount of heterogeneous data streams in real-time. The rest of the paper is organised as follows. Section 1.1 presents the proposed methods developed and used in the study including empirical data analytics, deep learning and data fusion. Section 2 discusses the application of the new method to heterogeneous data; the results obtained and makes an analysis. Finally, the last section concludes the paper and describes the direction of the future work.

1.1 Proposed methods

Heterogeneous data include several data forms or types such as qualitative or quantitative, structured or unstructured, primary or secondary. In reality, most situations or applications will produce heterogeneous data. For example, when a crime takes place; the evidence can be in the form of various types of data. The combination of all the data types can produce rich information insights. Nowadays, data become 'big' because they are generated every second, every day and the processing part is time-consuming and tedious. Also, each different data type has to undergo the processing differently, based on different data features. Feature extraction, thus, needs to be carried out per modality to select the important features. This phase is being called pre-processing and is different for image, video, text, and signals. After processing data, anomaly/outliers can be detected, and outliers should be removed to prevent error in the results (Liu et al. 2004). However, in some situations, for example, in the forensic applications, anomalies/outliers can represent significant information.

The traditional approach detecting anomaly is by using statistical methods which are based on a frequentistic (Chebyshev inequality) distribution approach to probabilities and threshold values. The thresholds are based on the distribution of data normal (Gaussian) or arbitrary (Chebyshev inequality). As studied in Angelov (2014), the traditional statistical approaches have the following disadvantages:

(a) they require *prior* knowledge of the distribution and, usually, an assumption of a (normal/Gaussian) distribution,
(b) using 3σ threshold values sometimes cannot detect the obvious anomaly/outliers,
(c) they require a large subset of data, and
(d) the information is blurred if they are compared with the average, instead of comparing in pairs of data samples.

These disadvantages can be avoided by using the recently introduced new computational framework called Empirical Data Analytics (EDA) (Angelov et al. 2017; Angelov 2015; Angelov et al. 2016). EDA does not require any *prior* assumptions and is based on the empirical obser-

vations in the data space. One of the main quantities in EDA is the standardised eccentricity (Angelov et al. 2017, 2016) which can detect anomaly without *prior* assumptions. Standardised eccentricity is very sensitive to abnormal data. It provided a measurement of the aggregation of the data item and is related to the tail of the data distribution.

Anomalies can be detected in each type of heterogeneous data individually. All of these anomalies can be combined by using data fusion techniques to create more significant analysis, result or information. Data fusion provides an integration of data from numerous sources and has been used to enhance information or decision-making (Castanedo 2013). Data fusion techniques have been widely applied in multi-sensor environments (Khaleghi et al. 2013). The motivations of using data fusion are various (Lahat et al. 2015). One of them is to improve the decision-making. Data fusion is a challenging task (Lahat et al. 2015). One of the challenges is working with heterogeneous datasets when advantages of each dataset are exploited maximally and the disadvantages suppressed.

In this paper, we propose a data fusion technique to detect anomalous behaviour by combining different data modalities. We demonstrate three types of data, namely signals (GPS data), financial (credit card and loyalty card data) and image data, as shown in Fig. 1. All the data have different values, types and dimensions. As a first step, we applied features extraction, and in regard to the features, we applied anomaly detection as a second phase. Anomaly detection is applied by using EDA, which is a standardised eccentricity and based on unsupervised techniques. This is only applied to GPS and financial data, while image data has been processed by using deep learning techniques. The next phase is data fusion through the combination of all the data types by using the approach of data fusion. In this technique, the first step is normalisation where all the data has to be transformed into a value between 0 and 1 to make the data comparable. Then, all the data can be integrated which later can form a contextual link with event sequences and storylines. Next, the results of fusing the data can be analysed more efficiently since the amount of data is significantly smaller, more organised and represent human-intelligible information in the form of graphs, tables and visualisations. The data fusion techniques can also represent the expert knowledge which later can be used to formulate the sequence of events. A human expert will need to verify the analysis in the final step and produce the final decision based on highly intelligible information, which is much smaller amount than the raw data. The significance of the data fusion approach is to assist the human expert in getting the right decision with a minimum time, nevertheless using a much larger amount of more complex heterogeneous data as a starting point.

1.2 Empirical data analytics

Empirical Data Analytics (EDA) is a new data analysis technique (Angelov et al. 2017; Angelov 2015; Angelov et al. 2016) which can be applied to clustering, classification, prediction and anomaly detection. This technique is a foundation of data analytics for statistic and streaming data analysis. Instead of analysing data using classical probability theory and analytics, EDA can be applied, respectively. EDA is more sensitive and flexible when detecting anomaly as compared to the traditional probability approach because it does take into account explicitly all the data (not only the mean). The anomaly can be detected without making a *prior* assumption about the data distribution. There are four main quantities which represent the properties of data in EDA (Angelov et al. 2017; Angelov 2015; Angelov et al. 2016). These include cumulative proximity, standardised eccentricity, data density and typicality. We applied standardised eccentricity to detect an anomaly in the datasets which can analyse local anomalies as well as global anomalies (Angelov et al. 2017). The standardised eccentricity, $\varepsilon(x)$ of a data points, x in EDA, is defined as follows (Angelov 2014):

$$\varepsilon(x) = \frac{2N\left[\sum_{j=1}^{N} d^2(x, x_j)\right]}{\sum_{j=1}^{N}\sum_{i=1}^{N}\left[d^2(x_i, x_j)\right]} \tag{1}$$

The traditional method considers 3σ and assumes in the normal data distributions. σ is the standard deviation. In real-world applications such as in engineering, natural science, biomedical and human behaviour modelling, distribution of data is various and hard to define. These are collected real-world data and not theoretical *prior* assumptions. There is not more than $1/n^2$th of the data sample that is abnormal (Angelov 2014). Normally, people prefer to apply 3σ or 6σ in a way to detect anomaly which guarantees no more than 1/9th (or $\sim$11%) if apply 3σ or no more than 1/36th (or $\sim$3%) if apply 6σ (Angelov et al. 2017). In regard to Chebyshev theorem (Saw et al. 1984), if the data sample has a high value of the standardised eccentricity ($\varepsilon(x) > n^2 + 1$) (Angelov et al. 2017) where n denotes the number of sigmas, then the anomaly is suspected. If we apply data streams, then it is efficient to calculate σ and ε recursively because the data will not be stored in the memory and it is computationally much more efficient. Standardised eccentricity can be defined as (Angelov 2015; Angelov et al. 2016):

$$\varepsilon(x) = 1 + \frac{\| \mu - x \|^2}{X - \| \mu \|^2} \tag{2}$$

where

$$X \leftarrow \frac{N-1}{N}X + \frac{1}{N}\| x_N \|, X \leftarrow \| x_1 \|^2 \tag{3}$$

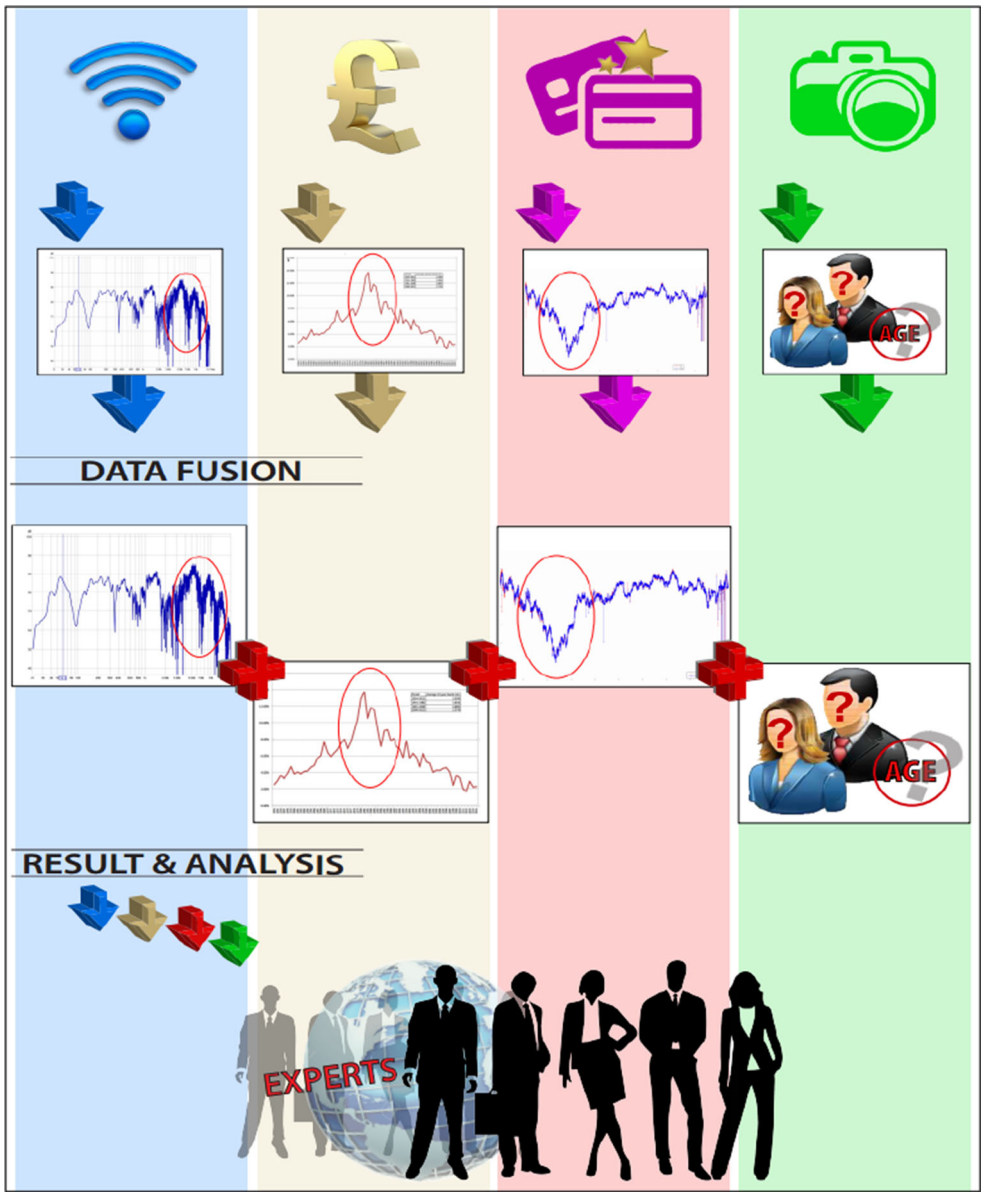


Fig. 1 Proposed method

$$\mu \leftarrow \frac{N-1}{N}\mu + \frac{1}{N}x_N, \quad \mu \leftarrow x_i \tag{4}$$

1.3 Deep learning

Deep learning is an advanced machine learning method and a computational model that consists of multiple processing layers to produce learning representations of data with multiple levels of abstraction. This model allows the computer to build complex concepts and improve and advance the image processing, speech recognition, and object recognition (Lecun et al. 2015). An example of deep learning model is the deep feedforward network composed of many simpler mathematical functions to map a set of input to output values. Convolutional neural networks (CNNs) is a type of deep learning based on feedforward network which is much easier to train and have better generalisation than networks with full connectivity between adjacent layers (Lecun et al. 2015).

In this paper, transfer learning and deep convolutional neural networks (CNNs) are applied to extract features from images of possible suspects. A vast amount of data and high computational resources are required to train a new deep learning model "from scratch". In some cases, the task is more challenging and requires training of few days. A large dataset is used to train CNNs which can then be fine-tuned for a specific task even in a different domain. This concept is called transfer learning or domain adaptation recently (Lecun et al. 2015) and includes solving a new task by applying the

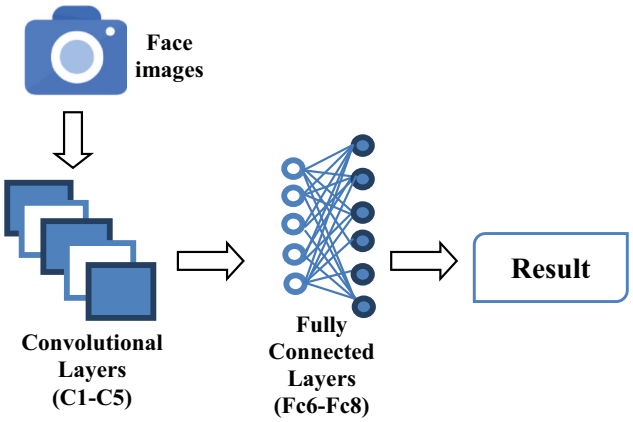


Fig. 2 AlexNet structure

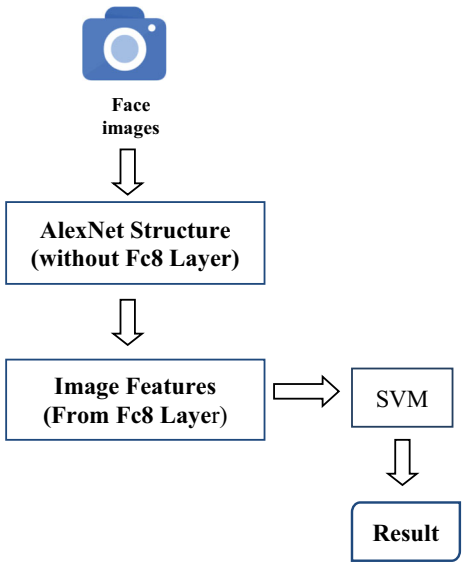


Fig. 3 Pre-trained net application structure

previously learnt knowledge. Transfer learning involves two major techniques, namely (1) update the weights on the new training dataset and keep the original pre-trained network, and (2) feature extraction and representation which will use pre-trained network then classify by a general classifier such as SVM (Zeiler and Fergus 2014). CNNs are likely to overfit with a small dataset. Therefore, transfer learning is very suitable for model training with the limited size of the dataset.

Figure 2 shows an AlexNet structure, and the input of AlexNet is an RGB image with 227 × 227 pixels and has five convolutional layers (from C1 to C5) and three fully connected layers (from Fc6 to Fc8). The activation function is Rectified Linear Unit (ReLU) (Krizhevsky et al. 2012). ReLU is defined as:

$$\mathrm{ReLU}(x) = \max(x, 0) \tag{5}$$

where x denotes the data points. There are 60 million parameters in this architecture. Due to a large number of parameters and the few thousand of training images, it is time-consuming. Therefore, transfer learning is very convenient. Figure 3 shows the combination of AlexNet and SVM classifiers to classify new images. There are three steps in the algorithm of this proposed solution, namely (Wang et al. 2017):

1. All the face images are normalised into a fixed size.
2. AlexNet will extract the image features using pre-trained deep learning network.
3. SVM will classify the images using features extracted at step 2.

1.4 Data fusion

Data fusion is applied to the heterogeneous data based on the newly proposed overall framework for anomaly detection from heterogeneous data streams/sets. There are five types of data with three different modalities and having different dimensionality. All the data cannot be simply combined and integrated. Therefore, we introduce a data fusion technique which first analyses the abnormality in each data type separately and determines the degree of suspicious between 0 and 1 and sums up all the degrees of suspicion data afterwards. There are four steps to transform all the data. Firstly, estimate the credit card data eccentricity. Secondly, find the difference between credit card and loyalty card data. Thirdly, locate the distance of the suspicious person's car to the store where the credit or loyalty cards were used, and, lastly, for images, we apply the accuracy of the classification based on their gender and age as a degree of confidence. After that, all the data are summed up together and divided by the number of data to get a normalised value of the level/degree of suspicion. The degree of suspicion based on the credit card data λ_k^{cc} is calculated based on the standardised eccentricity result. Eccentricity has been applied before to find an anomaly in credit card data. From eccentricity results, we calculate the degree of suspicion of the credit card data as follows:

$$\lambda_k^{cc} = 1 - \frac{1}{\varepsilon_k} \tag{6}$$

k denotes the data points. If the value of λ_k^{cc} is more than 0.9615, then we consider the specific use of the credit card to be suspicious, and if it is less than 0.9615, then it is considered normal. The abnormal value can be determined by the value of n we set in the dataset when calculating the ε. For example in these data, we consider the value of $\varepsilon = 26$ which corresponds to 5σ according to the Chebyshev inequality (Mohd

Ali et al. 2016). Thus, $\lambda_k^{cc} = 1 - \frac{1}{26} = 0.9615$ is considered as the threshold.

$$0 \leq \lambda_k^{cc} < 1 \tag{7}$$

Then, we calculate the disagreement between the credit card and loyalty card data, λ_k^{dis}. Firstly, the difference between the credit and loyalty card data is calculated. All the values are matched based on the timeline. After that, the difference is calculated to get the absolute disagreement. If the credit card and the loyalty card have the same value, then the value of δ is 0, which is considered as not suspicious:

$$\delta_k = \parallel cc_k - lc_k \parallel \tag{8}$$

$$\lambda_k^{dis} = 1 - \frac{1}{1 + \delta_k^2} \tag{9}$$

After that, the degree of suspicion based on the distance between persons, car and the location of the store λ_k^{loc} is calculated. We begin with calculating the distance d for every person's car from every location of a store. The calculation of λ_k^{loc} is conducted as follows:

$$\lambda_k^{loc} = e^{-\frac{d_k^2}{2\sigma_k^2}} \tag{10}$$

The value of sigma (σ) can be set based on the distance between a car park and the store location. We calculated the standard deviation for the distance from every store location to the car park for all trips every day and found the average. Based on this, we determined $\sigma = 555$ m. According to Van Der Waerden and Timmermans (2017), the normal distance between a car park and a store is between 50 and 700 m. The value we determined (555 m) is well within these limits. If the distance d between the person's car and the store's location is more than 555 m, then we consider the degree of suspicion to be high.

Data containing face images are used applied to identify the gender and the age of the person who used the card. We recognise the gender and age by using pre-trained deep CNN learning and SVM classifier. As a result, we use the classification accuracy to get the value of the degree of suspicion (if the gender or age do not match the true ones) based on the images. If both the age and the gender based on the images are the same as the true ones, then we calculate as below:

$$\lambda_k^{gen} \text{ or } \lambda_k^{age} = 1 - \text{classification accuracy} \tag{11}$$

To accommodate the uncertainty due to the classifier error. If the images have a difference of the gender or age, then we only take the classification accuracy value for the same reason.

$$\lambda_k^{gen} \text{ or } \lambda_k^{age} = \text{classification accuracy} \tag{12}$$

The final step is a fusion of all partial degrees of suspicion based on the partial data, λ_k^{total}. We sum all the values λ_k^i and multiply by weights. The weights can be set by experts based on the importance of the specific type of data. By default, weights can be set to 1:

$$w_i = 1, \quad i = \forall$$

Then, the sum of all values are divided into the number of types of data, N to normalise . λ_k^{total} is defined as follows:

$$\lambda_k^{total} = \frac{\sum_{i=1}^{N} w_i \lambda_i}{\sum_{i=1} w_i} \tag{13}$$

$$\lambda_k^{total} = \frac{w^{cc}\lambda_k^{cc} + w^{dis}\lambda_k^{dis} + w^{loc}\lambda_k^{loc} + w^{gen}\lambda_k^{gen} + w^{age}\lambda_k^{age}}{\sum_{i=1} w_i} \tag{14}$$

The data fusion technique we used represents a weighted average. The main novelty is that we consider heterogeneous data, and each variable represents a different type of data. All the data are already transformed into values between 0 and 1 to make them comparable. Then, we rank into a descending order of the λ_k^{total}. The highest value will be the most suspicious data. The lowest may be the least suspicious data. A human expert can determine whether a case is suspicious or non- suspicious by selecting a threshold value. If there is a need to have a line between suspicious and non-suspicious, then 3 sigma or Chebyshev inequality may apply to the result. At the end of the day, we do not intend (and hardly someone will allow or be happy with) a fully automatic system that resolves investigations. The aim is rather to simplify the role of the humans, to facilitate, so that they can focus on a small number of cases to be looked in more detail. The proposed approach does simplify the way of processing such huge amount of data, and later this method can assist the human expert in their investigation and making the final decisions.

2 Case study

In this paper, we consider as a test of the proposed new methodology data taken from the publicly available IEEE Visual Analytics Science and Technology (VAST) Challenge 2014 (VAST Challenge 2014). These data are a popular benchmark, and we are going to use these data as a proof of concept without limiting the generality of overall methodology. This challenge is about finding out suspicious patterns of behaviour that may be related to the missing staff members in the fictions GASTech Company. There are four datasets, namely credit card and loyalty card transactions, GPS data and car assignment (Table 1). Credit card data have 1492 data

Table 1 Description of dataset

Datasets	No of data points
Credit card	1492
Loyalty card	1393
GPS	685170
Car Assignment	45

Table 2 GAFace dataset

	Below 40 years old	Above 40 years old	Total
Male	82	84	166
Female	78	66	144
Total	160	150	310

points (transactions), while loyalty card data have 1393 data points. GPS data are the biggest dataset because these data represent the ability to track every movement of every staff member with 1Hz frequency. There are 685,170 data points in the GPS dataset. Information about a staff member and company car is stored in the car assignment dataset. Some of the employees in GASTech are given a company car which has installed with a geospatial tracking device. If the staff member does not have a company car, they can use company trucks for a business purpose. As the truck drivers' data do not have an ID, we removed the data related to truck drivers. The dataset is taken from 6 January to 19 January 2014.

In VAST Challenge 2014, there is no image dataset. To proof the concept of analysing heterogeneous data, we add one more dataset which includes image data. Finding publicly available dataset with the application of heterogeneous data is hard. Therefore, we build a dataset called GAFace which has 310 face images. GAFace is a face image dataset which was collected from Google Image, consisting of images of celebrities and politician in Malaysia. Data were collected from December 2016 to February 2017. There are 166 face images of male gender and 144 face images of female people. We divided the data into two age groups: below 40 years old (160 images), and above 40 years old (150 images) (Table 2).

2.1 Feature extractions

(a) Credit card and loyalty card data

Credit card and loyalty card data are a financial data about the money spent by every staff members. When a staff member uses his/her credit card, they usually get points on their loyalty card as well. There is a case where the staff members forget to bring their loyalty card, and they did not get award points. In some cases, the point in the loyalty card can be redeemed to buy something. In these datasets, credit card and loyalty card have five attributes, namely timestamp, location, money spend, first name, and last name. We only extract the money spent from this dataset. C_i is the money spent according to the credit card, and L_i is money spent according to the loyalty card.

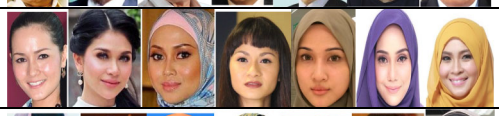
$$x_i = [C_i, L_i] \tag{15}$$

where i denotes the transaction index. From the credit card and loyalty card data, new features can be created; for example, total spending per person, total spending per person and per day, total spending per location and total spending per location per day. If we sum up every spending in credit card and loyalty card for every staff member, then we can create a new feature of "total spending per person". There are 35 staff members (excluding the truck drivers), and we can have 35 data points about the total spending of each staff member. These data can give information about who is spending more or less. Next, we can sum up the daily expenditure for every staff member. Then, we can extract features of "total spending per person per day". Regarding every location, we can create features of the total amount of spending per location. Therefore, we can get information of which popular location and staff member do spend more and how would that differ from others. Suspicious spending behaviour can be shown in this set of data. This will be easy to see when the suspicious behaviour takes place and the location is being determined. The money spent is normalised between 0 and 1 to make the data comparable. However, using one type of data reveals only a part of the complex picture and leaves too much uncertainty.

(b) GPS data

GPS data are generated from the geospatial software which is installed in the company car. There are 35 staff members who received a company car. Therefore, the information about the trajectories of all the 35 staff members can be easily tracked. There are four attributes in this dataset which are timestamps, ID, longitude and latitude. GPS coordinates which consist of longitude and latitude are the most important features in this dataset. Longitude and latitude can produce information about the mobility of staff members such as the projection of the trajectory, average speed, ratio of the trajectory angle, and distance. The formula to calculate all of these features can be found in Mohd Ali et al. (2016). A projection of the trajectory can be calculated from the trajectory of the staff members from starting point to the ending point. By using this feature, trajectory projection per axis can be separated. The average speed can also be calculated to differentiate every vehicle. Distance can also have to be calculated to show the length of trajectory from start point to end point. Lastly, we

Fig. 4 Sample of GAFace dataset

Gender	Images	Age
Male		<40 years old
		>=40 years old
Female		<40 years old
		>=40 years old

calculated the ratio of the trajectory change angle. The angle is calculated to find the sharpness of the turning point (less than 90^o or more than 90^o). Normally, if people are moving to some destination, they will go straight and will revert sometimes but not often. Therefore, the angle may reach up to 90^o but, if they turn more than 90^o, then we need to consider that as abnormal.

The trajectory angle and ratio are calculated as follows (Mohd Ali et al. 2016):

$$\theta_{i,j} = \arctan(b_{i,j+1} - b_{i,1}, a_{i,j+1} - a_{i,1}) \tag{16}$$

$$\text{IF } (\theta_{i,j} < 90^\circ) \text{ THEN } (R_\text{i} \leftarrow R_\text{i} + 0) \text{ ELSE } (R_\text{i} \leftarrow R_\text{i} + 1) \tag{17}$$

$$\bar{R} = \frac{R_\text{i}}{N_\text{i}} \tag{18}$$

where $a_{i,j}$ denotes the latitude and $b_{i,j}$ denotes the longitude, R_i denotes the time the angle exceeds 90° during a single trajectory. If the angle is less than 90°, then it will give the number of normal values, $R_\text{i} \leftarrow R_\text{i} + 0$, else the number of abnormal values, $R_\text{i} \leftarrow R_\text{i} + 1$. N_i is a duration (in seconds) of the travel on a given trajectory.

(c) Image Data (faces)

Image data have different features and dimensions from the other data. There are many steps to pre-processing images such as image resampling, greyscale, segmentation, and noise removal. In this case, we use GAFace image dataset. The face image dataset is separated into two classification tasks which are "gender" and "age". We classify the whole dataset by using age classifier (below 40 years old and above 40 years old) and gender classifier (male and female). Face image dataset is divided into two sets, randomly for training (80% of the data) and testing (20% of the data). The utilisation of the pre-trained deep learning network is important to extract features from the images. All the images are resized to 227 × 227 pixels size. Then, the resized images are being sent to an AlexNet. All the features are obtained after running the pre-trained network. Each of image features vectors has 4096 dimensions (Fig. 4).

2.2 Result and analysis

The proposed EDA-based anomaly detection approach has been applied to a set of a credit card, loyalty card, and GPS datasets (Mohd Ali et al. 2016). Anomalous behaviour has been detected in the credit card spending pattern. Figure 5a presents the standardised eccentricity of credit card data and shows that the majority of the eccentricity values are between 1 and 2, which is a very low eccentricity. Figure 5b shows that the highest eccentricity value is 1400 ($> 35\sigma$). The results indicate that only one noticeable anomaly in these data has been detected. These correspond to one of the staff members spending more than $10,000 in one single transaction, and the value is very suspicious compared to the average of the spending for every other staff member including the other transactions of the same member of staff. From the data shown in Fig. 6, one can see that the value of the credit card spending and loyalty card is not the same in every store. However, most of the transactions have almost the same value for both the credit card and loyalty card, except stores numbers 1, 4, 5, 11, 12, 14, 15 and 17. In this case, when people spend $1, it is equivalent to 1 point. There are also cases, when some stores may offer more points on a certain sale items.

Store No. 11 shows that the credit card spending is too high compared to the value of the loyalty card and overall. Normally, the more people spend their money, the more they will get loyalty card points. Therefore, we made a conclusion that the person who spent so much money is not bringing his/her loyalty card together during the purchase. Then, we compared the data from GPS to check the trajectory of the individual who has been suspected to have anomalous behaviour. There are four features derived from

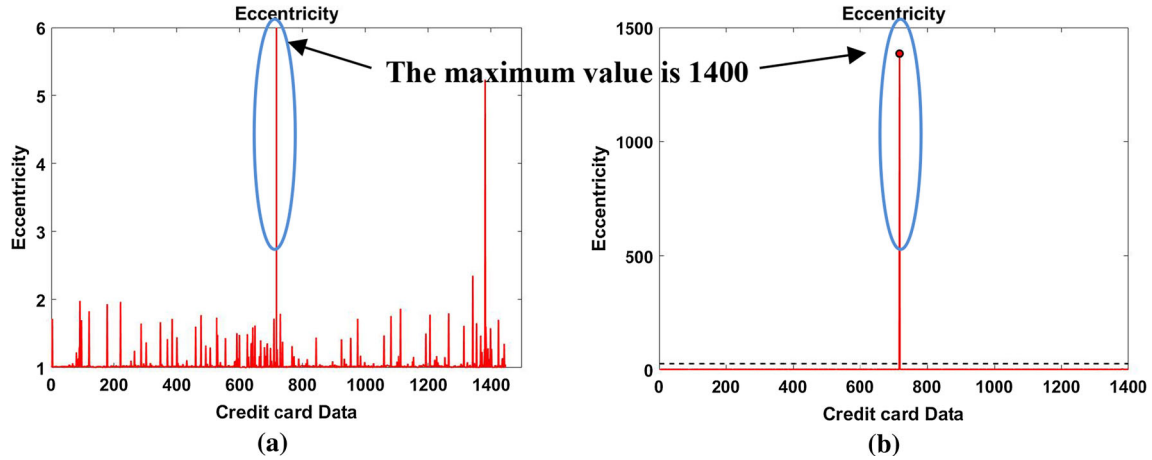


Fig. 5 Standardised eccentricity of credit card data (Wang et al. 2017)

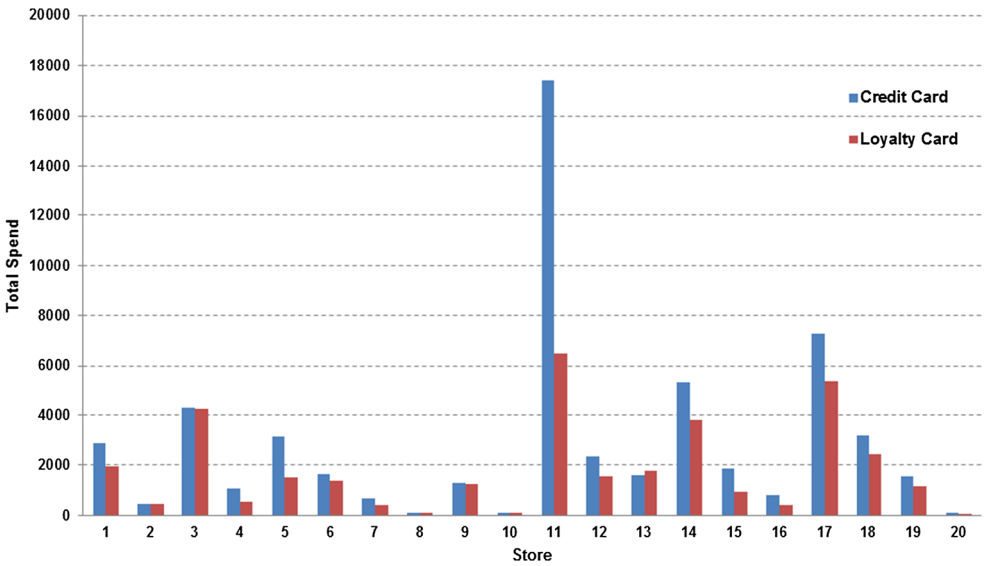


Fig. 6 Comparison between the usage of credit card and loyalty card in different locations (Mohd Ali and Angelov 2017)

the GPS data. One of the significant results is the ratio of the trajectory turn. Figure 7 provides the difference between the normal and abnormal trajectory pattern. There is only one staff member who is having an abnormal trajectory, but we consider the staff member as normal since his abnormal trajectory pattern shows daily consistency in any places he/she went, and we can make a conclusion that this is his/her routine travel pattern. After that, we check the GPS data on the day the suspicious transaction (the large purchase for $10,000 when the loyalty card was not used) and the time where the credit card spending is the highest amount. Figure 8 shows that during the day and time where the credit card is being charged for $10,000, the staff member is not in that location.

This further increases our suspicious, and we want to investigate further. Then, we become eager to prove who is the real person behind the transaction of such high amount of money. We check whether there is any surveillance camera

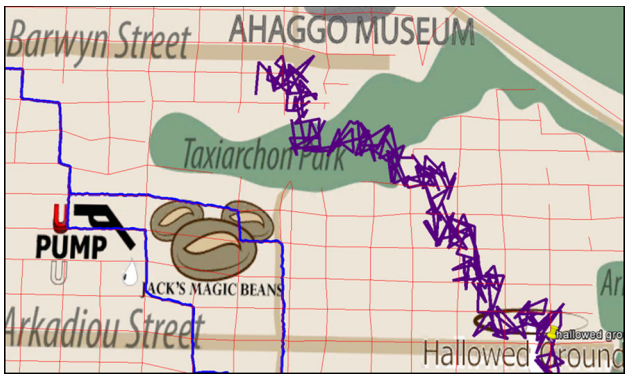


Fig. 7 Comparison between normal (blue line) and abnormal (purple line) trajectory pattern (Mohd Ali and Angelov 2017) (colour figure online)

at the location. We assume that security system collects the GAFace image dataset of faces, and we use it for recognition

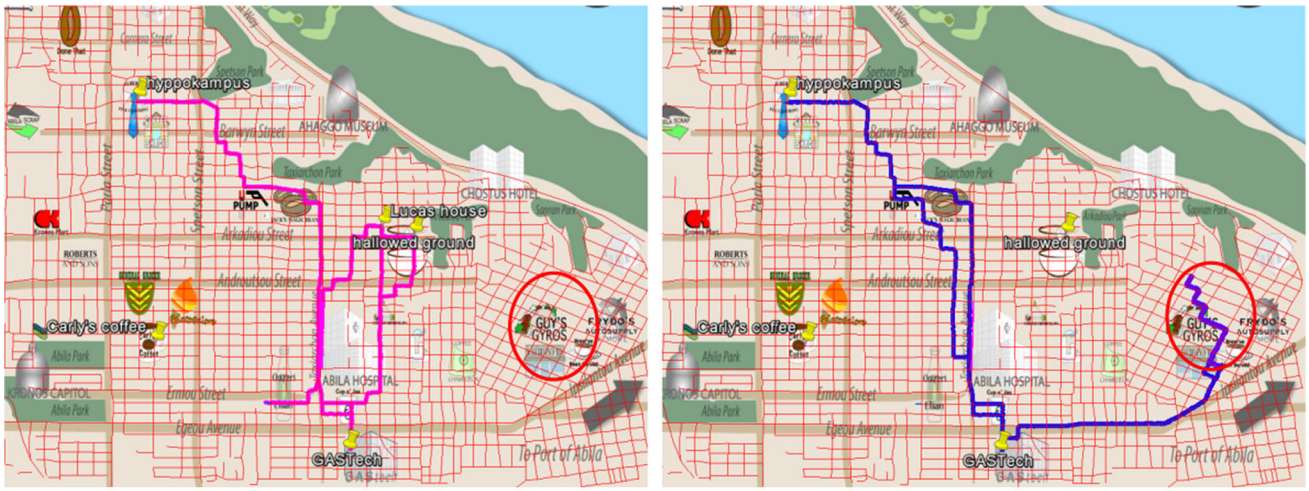


Fig. 8 GPS data of the car of the suspicious person (**a**) and a new person (**b**) (Mohd Ali and Angelov 2017)

Table 3 Classification results

Age	Gender
80.17%	90.33%

as to be applied in this scenario. We applied pre-trained deep neural network to extract the features and classify them by using SVM (Wang et al. 2017). Table 3 compared the accuracy results of age and gender classification. This result will be used by the proposed data fusion technique to show the integration of the different data types.

Data fusion is applied to combine all types and modalities of datasets. Before all the data being combined, there are several steps to be applied to ensure all the data are comparable and have the same range, [0,1]. Firstly, we calculate the degree of suspicion in credit card data by using Eq. (6). Secondly, we calculate the disagreement between credit card and loyalty card data. Thirdly, we calculated the degree of suspicion based on the distance between person's car and store location. Fourthly, we used the classification accuracy for face image data. Finally, we sum up all the data and divide by the number of datasets. After getting the fusion for every data, then, we sort out the result in ascending order to see which data are most suspicious. Table 4 shows the result of data fusion. The first row of data is the suspicious data where the degree of suspicion using a credit card is 0.99999896, degree of suspicion based on the credit card disagreement is 0.999298; degree of suspicion based on the gender of

Table 4 Results of data fusion

	Degree of suspicion					
No.	Credit card	Loyalty card	Age	Gender	Distance	Total
1	0.99999896	0.999298367	0.8017	0.0967	0.99897099	0.779334
2	0.55392285	0.006760843	0.1983	0.0967	0.9999928	0.371135
3	0.00038828	0.019266556	0.8017	0.0967	0.87355366	0.358322
4	0.69515606	0.627598032	0.1983	0.0967	0.08780634	0.341112
5	0.03434203	0.257665979	0.1983	0.0967	0.9999987	0.317401
6	0.27873867	0.004701328	0.1983	0.0967	0.99987384	0.315663
7	0.44699741	0.455473364	0.1983	0.0967	0.09044038	0.257582
8	0.59435563	0.006760843	0.1983	0.0967	0.08769619	0.196763
9	0.2594477	0.314689961	0.1983	0.0967	0.0000294	0.173833
10	0.32948267	0.024391952	0.1983	0.0967	0.0000499	0.129785
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1400	0.000190569	4.74405E−05	0.1983	0.0967	0.00000151	0.059048

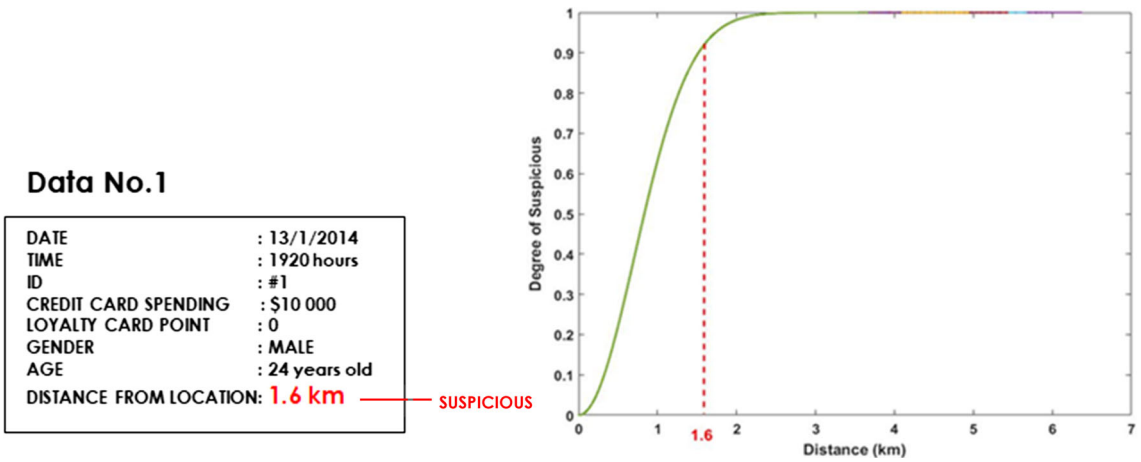


Fig. 9 Details of data No. 1

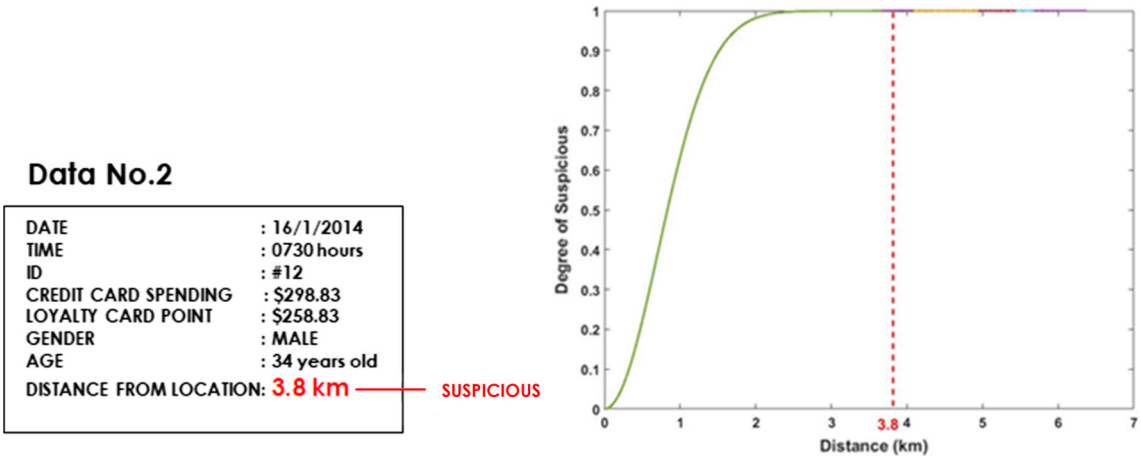


Fig. 10 Details of data No. 2

the suspected person taking into account the uncertainty due to gender accuracy is 0.0967, same for the age accuracy is 0.8017, and degree of suspicion in terms of distance between the store location and the person's car is 0.99897099. The overall data fusion result is 0.779334 which is the highest for any transaction made. From this result, the original data shows that on 13 January 2014 at 19.20pm, the suspect is spending $10 000, no credit card point recorded, the gender is male, age is 24, and the distance between the person's car and the store location is 1.6 km (see Fig. 9). The distance reveals that this person is not in the location when the card is being used. Then, we examine the day and the time at the location, and it shows that there is another person at the location. We suspect this person as to be the one who misused the credit card. The overall degree of suspicion is just over 78% which is somewhat less than 100% mostly because the gender of the person who used it and was on the surveillance cameras is the same as the real owner.

The result from the second row in Table 4 shows an anomaly in the credit card, loyalty card and distance. This person spent $298.83, and his loyalty card is only 258.83 point which is 40 points less than the total spent. This person is male and 34 years old. His location to the shop can be considered too far which is 3.8 km (see Fig. 10). Then, result no.3 shows the same person who is similar to the result no.1. The anomaly was detected based on the distance from his location to the store which is 1.58 km (see Fig. 11). The amount that this person used with the credit card is $55.25, which is not suspicious as compared to the first result. The same suspect is also detected in these data where the distance between him and the location is close.

Another example of analysis: if the gender of the person using the card is different, then the result will significantly change. From this case study, for most of the cases, the gender matches the gender of the owner. Therefore, we want to demonstrate that there will be a different result when the gender is different. Table 5 shows a case when in 4 cases the gender of the user is not the same as the one of the owner. Data line 1 shows an increase in the total degree of suspicion

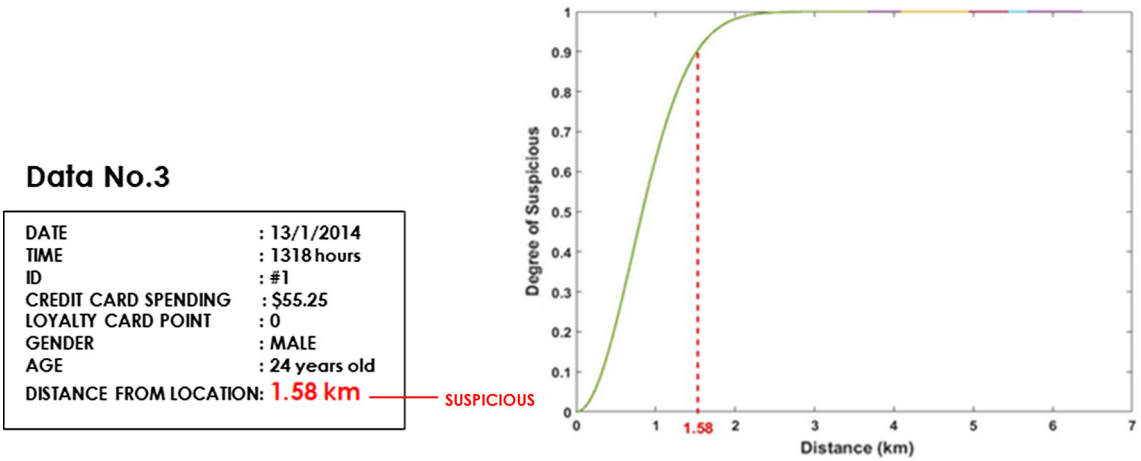


Fig. 11 Details of data No. 3

Table 5 Example with several cases of different gender

	Degree of suspicion						
No	Credit card	Loyalty card	Age	Gender	Distance	Total ($w_i = 1$)	Weighted total
1	0.99999896	0.999298367	0.8017	0.9033	0.99897099	0.940654	0.922063259
2	0.55392285	0.006760843	0.1983	0.9033	0.9999928	0.532455	0.73661143
3	0.00038828	0.019266556	0.8017	0.9033	0.87355366	0.519642	0.722932253
4	0.69515606	0.627598032	0.1983	0.9033	0.08780634	0.502432	0.666868667
5	0.03434203	0.257665979	0.1983	0.0967	0.9999987	0.317401	0.147601639
6	0.27873867	0.004701328	0.1983	0.0967	0.99987384	0.315663	0.183826492
7	0.44699741	0.455473364	0.1983	0.0967	0.09044038	0.257582	0.194545169
8	0.59435563	0.006760843	0.1983	0.0967	0.08769619	0.196763	0.201443977
9	0.2594477	0.314689961	0.1983	0.0967	0.0000294	0.173833	0.145475508
10	0.32948267	0.024391952	0.1983	0.0967	0.0000499	0.129785	0.144968628
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1400	0.000190569	4.74405E−05	0.1983	0.0967	0.00000151	0.059048	0.077891

after the data fusion from 0.779412 to 0.940732. Another example is in data no 2, 3 and 4, the total degree of suspicion after the data fusion rises from 0.371135 to 0.532455, 0.358322 to 0.519642 and 0.341112 to 0.502432, respectively. This shows that the difference in the gender can help in finding the suspicious cases. We also add another column for the weighted total. Weight can be added and adjusted by the human expert. They can set the weights based on the importance of the data. In this example, we assume that gender is most important because if the gender is different, then apparently it is suspicious. Therefore, we set the higher weight for gender which is 0.6. Then, for credit card and age data, weight is set to 0.2 and 0.1, respectively. For the other two variables, loyalty card and distance location, the weight is set to 0.05. After calculating the weighted degree of suspicion, the most suspicious case is still the same data like 1 for which the degree is 0.922063259. Then, the degree of suspicion for lines 2, 3 and 4 rises from 0.532455 to 0.73661143, 0.519642 to 0.722932253 and 0.502432 to 0.666868667, respectively.

Figure 12 presents the comparison of the number of data to be process and number of 10 most suspicious fused data. Before we get the most suspicious fused data to be investigated, there are many data to process, and it is time-consuming and tedious. Using our proposed method can help investigator or expert to simplify all the data and sort based on the level of suspicion. From these three analyses, it is obvious that every data item has been simplified and helpful for the investigator to shorten the time in investigating fraud cases or other crime or anomalous behaviour.

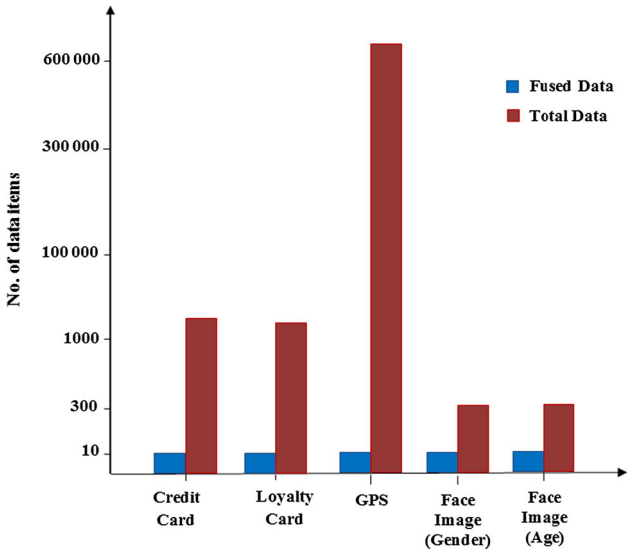


Fig. 12 Comparison between the amount of the original data and the amount of the fused data

3 Conclusion

In this paper, we proposed a data fusion technique to identify anomalous behaviour based on heterogeneous data. The illustrative example of using VAST Challenge data (VAST Challenge 2014) and GAFace dataset shows the application of data fusion technique on heterogeneous data. We demonstrate five types of data, namely a credit card, loyalty card, GPS, age, and gender based on face images. The anomaly is detected from the credit card data, loyalty card and GPS data. There is one person who is consistent in all the anomalies. We also remove these data, and then a new anomaly has been detected, but after we check the details of this person, he is the CEO of the company. Therefore, it is normal for him to spend more than the other staff members. Image data are processed by using pre-trained deep learning network, and then the classification is done using SVM. Anomaly result and image classification result are combined through the data fusion technique. The result from fusion technique is then ranked ascendingly to examine which is the suspicious data. After analysing the result, we discovered a new suspicious person. Location distance between the staff members and the store is far (1.6 km). We can conclude that he is not at the store when the credit card is charged. Analysis of this result can assist the human expert in simplifying their job and helping them in making a decision. In the future research, we plan to have a variety of data types such as text data or streaming data and applying to them the proposed data fusion technique.

Acknowledgements The first author would like to acknowledge the support from the Ministry of Education Malaysia and Universiti Teknologi MARA, Malaysia, for the study grant.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interests.

Ethical approval This article does not contain any studies with human participants or animals performed by the author.

References

Abdallah A, Maarof MA, Zainal A (2016) Fraud detection system: a survey. J Netw Comput Appl 68:90–113

Angelov P (2014) Anomaly detection based on eccentricity analysis. In: IEEE symposium on evolving and autonomous learning systems (EALS), 2014. pp 1–8

Angelov P (2015) Typicality distribution function—a new density—based data analytics tool. In: IJCNN 2015 international joint conference on neural networks. pp 1–8

Angelov P, Gu X, Kangin D (2017) Empirical data analytics. Int J Intell Syst 0:1–24

Angelov P, Xiaowei G, Kangin D, Principe J (2016) Empirical data analysis: a new tool for data analytics. In: IEEE international conference on systems, man, and cybernetics. pp 52–59

Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. WWW 2012—Session: information diffusion in social networks, April 16–20, 2012, Lyon, France. pp 519–528

Castanedo F (2013) A review of data fusion techniques. Sci World J 2013:1–9

Delgado B, Tahboub K, Delp EJ (2014) Automatic detection of abnormal human events on train platforms. In: IEEE National aerospace and electronics conference 2009:169–173

Dhar V (2013) Data science and prediction. Commun ACM 56(12):64–73

IDC (2013) Where in the world is storage: a look at byte density across the globe

Jyothsna V (2011) A review of anomaly based intrusion detection systems. Int J Comput Appl 28(7):975–8887

Keogh E, Lin J, Fu AW, Van Herle H (2006) Finding unusual medical time-series subsequences : algorithms and applications. IEEE Trans Inf Technol Biomed 10(3):429–439

Khaleghi B, Khamis A, Karray FO, Razavi SN (2013) Multisensor data fusion: a review of the state-of-the-art. Inf Fusion 14(1):28–44

Kim Y, Kogan A (2014) Development of an anomaly detection model for a bank's transitory account system. J Inf Syst 28(1):145–165

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1–9

Lahat D, Adali T, Jutten C (2015) Multimodal data fusion: an overview of methods, challenges, and prospects. IEEE Proc 103(9):1449–1477

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nat Int Wkly J Sci 521:436–444

Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32

Liu H, Shah S, Jiang W (2004) On-line outlier detection and data cleaning. Comput Chem Eng 28:1635–1647

Malekian D, Hashemi MR (2013) An adaptive profile based fraud detection framework for handling concept drift. In: 10th international ISC conference on information security and cryptology, ISCISC 2013, pp 1–6

Mohd Ali A, Angelov P (2017) Applying computational intelligence to community policing and forensic investigations. In: Community policing—a European perspective, pp 1–16

Mohd Ali A, Angelov P, Gu X (2016) Detecting anomalous behaviour using heterogeneous data. In: Contributions presented at the 16th UK workshop on computational intelligence advances in computational intelligence ssytems, Sept 7–9 2016. Lancaster, UK, pp 253–273

Pallotta G, Vespe M, Bryan K (2013) Framework for anomaly detection and route prediction. Entropy 15:2218–2245

Palmieri F, Fiore U, Castiglione A (2014) A distributed approach to network anomaly detection based on independent component analysis. Concurr Comput Pract Experience 26(5):1113–1129

Philip Chen CL, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci (Ny) 275:314–347

Pollet TV, van der Meij L (2016) To remove or not to remove: the impact of outlier handling on significance testing in testosterone data. Adapt Hum Behav Physiol 3(1):43–60

Salem O, Guerassimov A, Marcus A, Furht B (2013) Sensor fault and patient anomaly detection and classification in medical wireless sensor networks. In: IEEE ICC 2013—selected areas in communications symposium, pp 4373–4378

Saw JG, Yang MCK, Mo TSEC (1984) Chebyshev inequality with estimated mean and variance. Am Stat Assoc 38(2):130–132

Van Der Waerden P, Timmermans H (2017) Car drivers characteristics and the maximum walking distance between parking facility and final destination. J Transp Land Use 10(1):1–11

VAST Challenge 2014 (2014) (Online). www.vacommunity.org/VAST. Challenge 2014

Wang X, Mohd Ali A, Angelov P (2017) Gender and age classification of human faces for automatic detection of anomalous human behaviour. In: International conference on cybernetics (CYBCONF 2017), pp 1–6

Wu Y, Patterson A, Santos RDC, Vijaykumar NL (2014) Topology preserving mapping for maritime anomaly detection. Springer, Cham, pp 313–326

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision, pp 818–833

Zuech R, Khoshgoftaar TM, Wald R (2015) Intrusion detection and big heterogeneous data: a survey. J Big Data 2(1):1–41