IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 29, 2019, accepted June 25, 2019, date of publication July 3, 2019, date of current version August 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926518

Importance of Small Probability Events in Big Data: Information Measures, Applications, and Challenges

RUI SHE[1,2], SHANYUN LIU[1,2], SHUO WAN[1,2], KE XIONG[3,4], (Member, IEEE), AND PINGYI FAN[1,2], (Senior Member, IEEE)

[1]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[3]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China
[4]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Pingyi Fan (fpy@tsinghua.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771283, and in part by the National Innovation Research Group Project under Grant 61621091.

ABSTRACT In many applications (e.g., anomaly detection and security systems) of smart cities, rare events dominate the importance of the total information on big data collected by the Internet of Things (IoT). That is, it is pretty crucial to explore the valuable information associated with the rare events involved in minority subsets of the voluminous amounts of data. To do so, how to effectively measure the information with the importance of the small probability events from the perspective of information theory is a fundamental question. This paper first makes a survey of some theories and models with respect to importance measures and investigates the relationship between subjective or semantic importance and rare events in big data. Moreover, some applications for message processing and data analysis are discussed in the viewpoint of information measures. In addition, based on rare events detection, some open challenges related to information measures, such as smart cities, autonomous driving, and anomaly detection in the IoT, are introduced which can be considered as future research directions.

INDEX TERMS Information measure, rare events, big data analytics, information theory, IoT, smart cities, autonomous driving.

I. INTRODUCTION

It is predicted that by 2050, the urban population all over the world is going to be doubled, which will rise up to 6.7 million people. With the rapid growth in the amount of urban residents, cities bring in new opportunities, while many new challenges come up, including environmental deterioration, sanitation problem, traffic congestion and terrorist attacks. In order to figure out these problems so that citizens may enjoy a new daily life with security and convenience, Internet of Things (IoTs) has been emerging as an effective solution [1]–[5].

In IoTs, explosively increasing sensors and devices are deployed to sense and collect different types of data, e.g., states of moving cars, crossroads and subway tracks, which drive us into a ‘‘big data’’ era. In order to make things smart, massive data has to be mined to find useful information and knowledge. In this case, the key point lies in how to deal with the observed data and dig out the hidden valuable information [6]–[13]. To do so, a series of promising technologies have been put forward such as statistical learning, computer vision, signal processing and so on [14]–[20].

The associate editor coordinating the review of this manuscript and approving it for publication was Lu An.

A. IMPORTANCE OF RARE EVENTS WITH SMALL PROBABILITY

As a matter of fact, in some applications, the regular patterns of systems’ or users’ behaviors are required to be explored from common events that often occur, but for the other applications, the rare events attract more attention than those occurring with large probability. For example, in financial crime detection systems, only a few illegal identities causing financial frauds indeed catch our eyes [21], [22], which are more important from subjective consciousness.

Besides, in intrusion detection systems, only a few number of security alarms should be detected and handled [23]–[27].

So far, a lot of works have investigated networking intrusion and reliable communications to protect IoTs from being attacked [28]–[38], which show that the rare events should be focused on for their special value in IoTs. By resorting to IoTs or other monitoring devices [39], smart city is becoming a timing fashion in city planning, construction, management and operations [40]–[45]. In this case, the rare events observed from monitoring systems also contain more significant features in the numerous data, which can provide effective references for transportation management, city planning and public safety.

Due to the fact that anomalous events may be hidden in big data [46]–[50], it is significant to process rare events or the minorities in objective detection. With regard to the autonomous driving in highways, it is crucial to detect the unexpected moving obstacles over lanes (which can be viewed as rare events). It is reported that around 150 people die from road hazards in American traffic accidents every year [51], [52]. It is beneficial to develop autonomous driving cars based on anomalous objective detection in many aspects such as reducing traffic congestion and accidents, improving energy efficiency and ensuring transportation safety. Actually, there are some researches trying to design intelligent vehicle systems to avoid dangerous driving events with small probability [53]–[59].

In brief, rare events have special values in many newly rising fields such as IoTs, smart city, and autonomous driving. Actually, the approaches for small probability processing are investigated from many perspectives in big data era.

B. INFORMATION THEORY FOR RARE EVENTS

In the viewpoint of information theory, information measures could have a seat on the table of rare events processing in big data. According to conventional information theory, the uncertainty of probability distributions can be characterized by information measures such as Kolomogorov complexity, Shannon entropy, Renyi entropy, and mutual information. These measures are also applicable to the infrequent or abnormal events [60], [61]. By using information measures to analyze the complexity of the different classes in big data, rare events would be recognized and handled [62]. For instance, an objective function of distribution was proposed based on factorization to detect the subsets with smaller probability [63]. Additionally, as an effective information distance, the relative (or differential) entropy is also applied to outlier detection [64]. Although there are special scenarios where the above approaches can be used, it is evident that they just focus on the large probability elements or subsets to deal with rare events.

From the perspective of small probability elements, there are also some technologies in the framework of statistical mechanics, such as the large deviation approaches and the measure of concentration of rare events [65], [66]. In these cases, traditional information measures are explained and extended by aiming at minority subsets processing. These technologies could be also used in many applications such as secure lossy compression and anomaly detection [67], [68].

In the framework of data distribution processing, the information divergence as a kind of information measure is an intersection of information theory and big data analytics. In fact, information divergences can be adopted to measure the distance between two distributions with small probability elements. Currently, information divergences have been used in many applications involved with rare events such as faulty detection [69], key frame selection [70] and image recognition [71], [72]. Therefore, how to use information measures to cope with small probability events becomes more interesting.

C. THE MOTIVATIONS AND CONTRIBUTIONS

The purpose of this paper is to integrate the works on importance analysis of small probability events and clarify the relationship between small probability cases with more importance and information processing including the corresponding information measures and applications. Essentially, this paper is not a technical work but a survey to summarize some classical theories and approaches of information processing based on small probability events so that the related literature can be discovered in a logical and reasonable way.

As far as the contribution of this work is concerned, a theoretical framework with a common fundamental form of message importance measure is constructed to show the core idea of importance of small probability events and characterize its mathematical representation. Moreover, similar to Shannon entropy, an information processing architecture is proposed from the perspective of message importance to combine the message importance compression, transmission loss and receiver preprocessing, which may broaden the extension of conventional information theory. In this case, some novel source coding strategies and information distortion analysis are obtained in an information system based on the importance of small probability events. For big data analytics, some related technologies including measures estimation, dimension reduction and correlation analysis are also unified into an architecture of information system to process important small probability events. This provides a reasonable data processing procedure for the small probability events hidden in massive samples. Finally, some modern and challenging applications, such as smart cities, autonomous driving, and IoTs, may adopt the information measures based on the message importance as novel criterions or metrics for rare events detection. In this regard, we present some schemes with information measures for the corresponding applications.

D. ORGANIZATION

The organization of the rest parts of this paper is summarized as follows. In Section II, we analyze some theories and technologies of information measures in the scenarios where rare events have valuable sense. In Section III, we discuss some applications based on information measures for rare events, including information compression, transmission

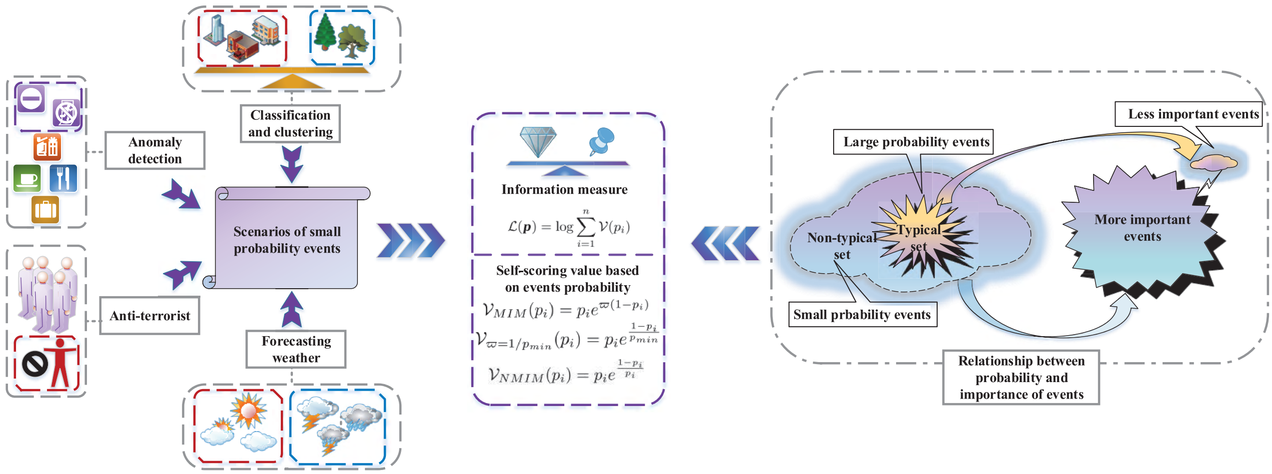


FIGURE 1. The interpretation of information importance measures focusing on rare events.

and preprocessing. Section IV first introduce some effective estimations of distributions and their functionals. Then, information coupling, directed information and some applications involved with rare events are introduced to reduce the dimension of big data and analyze the data causality or correlation from the perspective of information theory. In Section V, some challenging research directions for information measures are presented based on the rare events detection. At last, we conclude the paper in Section VI.

II. INFORMATION THEORIES AND TECHNOLOGIES FOR MEASURING RARE EVENTS

Information measures play important roles in not only traditional information theory but also numerous applications of big data, such as detection, classification and clustering [73], [74]. In fact, by facilitating the small probability elements, some information measures focusing on rare events are proposed to settle the big data problems such as anomaly detection, feature selection and pattern recognition [75]–[77]. In these cases, rare events can be extremely eye-catching, in good agreement with the fact that the vital part of the information attracts more attention than the perfect information. Consequently, in this paper, we merely focus on the cases where small probability events, referred to as rare events, contain importance of information.

To characterize the importance of rare events mathematically, Message Importance Measure (MIM) [78]–[82], fixed-parameter MIM [83] and NMIM (Non-parametric MIM) [84] are proposed, whose details are summarized in the Table 1. We also analyze the characteristics of these information measures and compare their similarities and differences as follows.

i) Intrinsic sense of the information measures: The common fundamental form for the information measures (including MIM, fixed-parameter MIM and NMIM) can be given by

$$\mathcal{L}(\boldsymbol{p}) = \log \sum_{i=1}^{n} \mathcal{V}(p_i), \tag{1}$$

where $\boldsymbol{p}$ is the given distribution which satisfies $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$, and the components $\mathcal{V}(p_i)$ of MIM, fixed-parameter MIM and NMIM are respectively given by

$$\mathcal{V}_{MIM}(p_i) = p_i e^{\varpi(1-p_i)}, \tag{2a}$$

$$\mathcal{V}_{\varpi=1/p_{min}}(p_i) = p_i e^{\frac{1-p_i}{p_{min}}}, \tag{2b}$$

$$\mathcal{V}_{NMIM}(p_i) = p_i e^{\frac{1-p_i}{p_i}}, \tag{2c}$$

where ϖ denotes the coefficient of importance. Actually, these values are just the same as the intuitive notion of importance value, which can be viewed as the invariant of system, referred to as *self-scoring value*. It implies that larger weights are allocated to the small probability events than those with large probability. Furthermore, Fig. 1 is shown to describe the above information measures visually. Specifically, by treating important events from the probabilistic viewpoint, the status of the atypical sets with small probability is highlighted, which can match many scenarios such as anomaly detection, anti-terrorist activities, forecasting abnormal weather, classification and clustering for binary events.

ii) Comparison of the information measures in the Bernoulli case: Here, the comparison of some different importance measures with respect to the Bernoulli distribution $(p, 1-p)$ is shown in Fig. 2.

It illustrates that the parameters of MIM can make great differences on the characterization for the Bernoulli distribution. While, the non-parametric MIM (namely NMIM) and the parametric MIM (namely fixed-parameter MIM) both have similar performance on measuring small probability elements. In brief, the details of comparison are listed as follows.

- Due to the fact that the MIM can be influenced by parameter ϖ, there is no worry about beyond the computing ability of computers.
- If the probability elements are small enough, MIM amplifies small probability not as greatly as NMIM and parametric MIM.

TABLE 1. Summary of information measures for rare events.

Metrics	Definition	Main Properties	Key Points
MIM with parameter ϖ [78]–[82]	$L(\boldsymbol{p},\varpi)=\log\sum_{i=1}^{n}p_i e^{\varpi(1-p_i)}$, with the parameter $\varpi \geq 0$. [Definition 1. in [78]]	• $L(\boldsymbol{p},\varpi)\geq 0$ with $\varpi\geq 0$. • $L(\boldsymbol{p},\varpi)\geq \varpi(1-\sum p_i^2)$. • $L(\alpha\boldsymbol{p}+(1-\alpha)\boldsymbol{q},\varpi)\geq L(\alpha\boldsymbol{p},\varpi)$ $+L((1-\alpha)\boldsymbol{q},\varpi), \alpha\in(0,1)$. • Event decomposition and merging: The MIM can be increased by dividing one event into two sub-events.	• Minority detection combined with the Bayes method. • Measure system states with small probability events. • Characterize users' preference distribution for recommendation.
The fixed-parameter MIM [83]	$L_j(\boldsymbol{p},\varpi_j)=\log\sum_{i=1}^{n}p_i e^{\varpi_j(1-p_i)}$, with the parameter selection given by $\varpi_j=F(p_j)$. [Eq. (3) in [83]]	• Principal component: p_j becomes the principal component in the MIM with $\varpi_j=1/p_j$. • $L_a(\boldsymbol{p},\varpi_a)<L_b(\boldsymbol{p},\varpi_b)$ if $p_a>p_b$. • $L_j(\boldsymbol{p},\varpi_j)\geq\log\left(\sum_{i=1}^{n}p_i e^{(-p_i)}\right)+1$.	• To focus on the probability p_j by using $\varpi_j=F(p_j)=1/p_j$. • Applied to minority subset detection. • The mean and variance can converge when samples $N_i\to\infty$.
Non-parametric MIM [84], [85]	$L_{non}(\boldsymbol{p})=\log\sum_{i=1}^{n}p_i e^{\frac{1}{p_i}-1}$, where $p_i>0$. [Definition 1. in [84]]	• $L_{non}(\boldsymbol{p})\geq 0$. • $L_{non}(\boldsymbol{p})\geq n-1$ • $L_{non}(\boldsymbol{p})+L_{non}(\boldsymbol{q})\leq L_{non}(\boldsymbol{pq})$ • Event decomposition and merging: The NMIM can be increased by dividing one event into two sub-events.	• Without the constraint of parameter selection. • Exponential operator rather than logarithm or polynomial operator. • Storage code design and transmission planning in wireless communication.

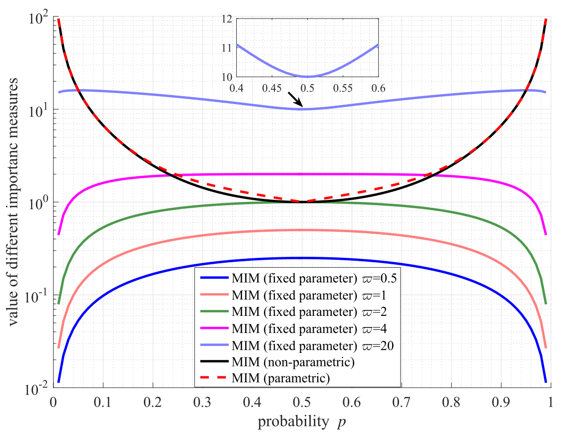


FIGURE 2. The comparison of different message importance measures with respect to the Bernoulli distribution (p, $1-p$).

- In the adjacent region of uniform distribution, the parametric MIM can perform better to amplify the smaller probability than NMIM.

III. APPLICATIONS IN MESSAGE PROCESSING

With respect to big data in IoTs, it is significant to design efficient strategies for message processing including information compression and transmission [86]. In particular, considering rapidly exploring data [87], we never need to store the whole data samples as before. Besides, since data traffic is exponentially increasing, it is a challenge for transmission resources (including links or networks) to carry so many data pockets [88]. Hence, the data processing techniques about lossy compression and transmission are investigated in many aspects [89]–[92]. In fact, information theory is a fundamental theory for data compression and transmission [93]. To be specific, it provides the optimal coding strategy and the tight bounds for the lossless and lossy compression [94]. Moreover, it also proposes information measures including relative entropy, Renyi divergence and f-divergence to guide information transmission and analysis [95], [96].

From the perspective of rare events, a message processing architecture based on message importance measure is presented as Fig. 3, whose details are listed as follows:

i) As for the information source, it is significant to maintain the rare events regarded as important message and lose some normal events. In this case, it is feasible to make use of the importance measures to design lossy compression schemes. To this end, the reconstruction error weighted by message importance can be minimized to achieve the lower bounds of code length.

ii) From the viewpoint of transmission for message importance, the core idea is that the receiver can gain more amount of information from the source while maintaining the affordable loss of message importance. In this case, the change of information measure focusing on rare events can be used to characterize the upper bound of information importance loss.

iii) In the information sink, it is possible to use some information divergences to distinguish two adjacent distributions containing rare elements. This can be regarded as a preprocessing for received data.

In terms of the specific analysis of the message processing architecture, three possible applications of information

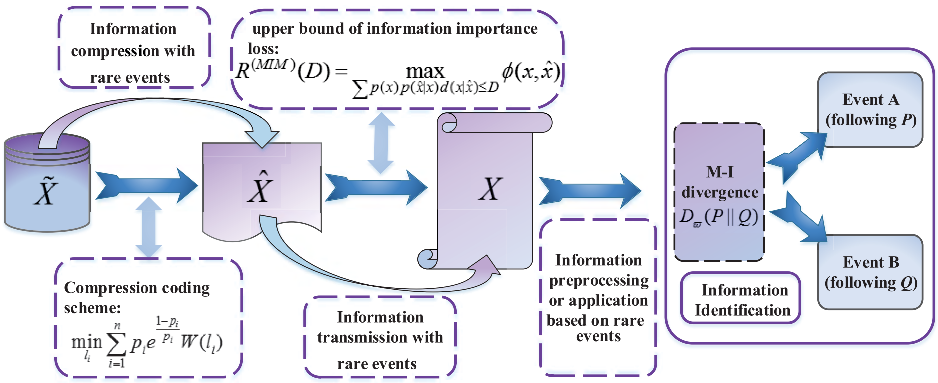


FIGURE 3. Message processing architecture from the viewpoint of rare events.

TABLE 2. Summary of message processing applications.

Work area	Description	Key Points
Information Compression [84]	Compression scheme based on NMIM: $\arg\min_{l_1,l_2,\ldots,l_n} \sum_{i=1}^{n} p_i e^{\frac{1-p_i}{p_i}} W(l_i)$ s.t. $C_n = \sum_{i=1}^{n} l_i$, if $l_i \le T$, $(i = 1, 2, \ldots, n)$,	• NMIM is regarded as a measure to weight the importance of code length. • Longer codewords are allocated to the rare events rather than the large probability ones. • Lower bounds of the code length in the sense of message importance. • $W(l_i)$ is selected by different expressions such as l_i^{-1} and γ^{-l_i} (γ is a constant).
Information Transmission [85]	NMIM loss distortion function with respect to distortion D: $R^{(MIM)}(D) = \max_{p(\hat{x}\|x)\in\Omega} \phi(X,\hat{X})$, $\Omega = \{p(\hat{x}\|x) : \sum_{(x,\hat{x})} p(x)p(\hat{x}\|x)d(x,\hat{x}) \le D\}$, $\phi(X,\hat{X}) = L_{non}(\mathbf{p}_x) - L_{non}(\mathbf{p}_{\hat{x}})$	• The upper bound of message importance loss caused by transmission distortion. • NMIM-loss-distortion viewpoint consisting of the message importance loss $\phi(X,\hat{X})$, the distortion D and the distribution of events.
Information Preprocessing [73]	The test method for outlier detection based on information divergence $\mathcal{F}(\cdot)$: $\mathcal{F}(\hat{\mathbf{p}}(\mathcal{X}^{(i)}); \hat{\mathbf{p}}_t) \to \begin{cases} \mathcal{F}(\hat{\mathbf{p}}_{t_i}; \hat{\mathbf{p}}_t), & \mathcal{X}^{(i)} \in \mathcal{M}_t \\ \mathcal{F}(\hat{\mathbf{q}}_{f_i}; \hat{\mathbf{p}}_t), & \mathcal{X}^{(i)} \in \mathcal{M}_f \end{cases}$,	• Distribution Estimations obtained by maximum likelihood estimator, k-nearest neighbor or Gaussian kernel estimator. • Message identification divergence given by Eq. (3) as $\mathcal{F}(\cdot)$ to detect the outlier sequences with small probability of occurrence.

measures are summarized in the Table 2, whose main details and interpretations are given as follows.

- *Information Compression:* Although standard compressions are proposed to reduce some redundant information in some degree [94], there still exists large size of data that contains some unimportant message. Further compression is considered to abandon the less vital message based on the probability of events, which may be achieved by using the compression scheme based on NMIM [84]. In this case, lower bounds of the code length l_i (with the limited total code length C_n) is obtained in the sense of message importance (based on the function of reconstruction error per unit importance, denoted by $W(l_i)$).
- *Information Transmission:* As far as big data is concerned, the dominant part of message with more importance is more favored rather than the redundant message. In the traditional information transmission, some distortions or errors may have more disastrous impacts on the important messages than worthless ones. For instance, based on this characteristic, the strategy of unequal error protection (UEP) codes has been proposed as a reliable transmission approach [97]–[99]. From a new viewpoint of rare events, data transmission with the constraint of message importance loss is discussed to guide the design of information transmission [85]. In particular, the upper bound of message importance loss $\phi(X, \hat{X})$ (based on NMIM operator $L_{non}(\cdot)$) is given when there exists a kind of distortion $d(x, \hat{x})$ (such as Hamming distortion) between a source X and a distortion source $\hat{X}$.
- *Information Preprocessing:* Considering the information preprocessing, information divergences play vital roles in discriminating different distributions (namely information identification). That is, the information

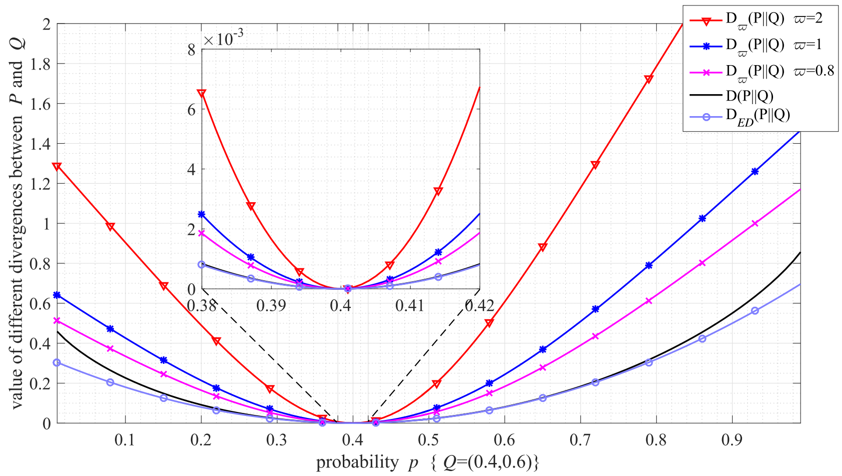


FIGURE 4. Comparison for different information divergences (between the probability distributions P and Q where $P = (p, 1-p)$ and $Q = (0.4, 0.6)$) including the MI divergence (with the parameter $\varpi = 2, 1, 0.8$), KL divergence and squared Euclidean distance.

divergence can be used as a test tool for outlier detection [73], [74], [100]. In particular, an information divergence between two distributions, denoted by $\mathcal{F}(\cdot)$, can classify the pending sample sequences $\mathcal{X}^{(i)}$ into the normal sequence set $\mathcal{M}_t$ or the outlier sequence set $\mathcal{M}_f$. In fact, the message identification (MI) divergence has its advantage on outlier detection [73], whose definition is given by

$$D_\varpi(\boldsymbol{p} \parallel \boldsymbol{q}) = \log \sum_{i=1}^{n} p_i e^{\left(\varpi \frac{p_i}{q_i}\right)} - \varpi, \tag{3}$$

where the adjustable coefficient ϖ is positive, as well as $\boldsymbol{p}$ and $\boldsymbol{q}$ are two finite probability distributions in the same support set. Here, we also take two Bernoulli distributions P and Q as examples to compare different information divergences shown in Fig. 4. It is illustrated that MI divergence described in the Eq. (3) is more sensitive to distinguishing two distributions than the Kullback-Leibler (KL) divergence and the squared Euclidean distance when the distribution P is closed to the distribution Q [73].

Remark 1: i) For information compression: As for the data compression based on information measures for rare events, the common core idea is that the code length mainly depends on the message importance of events. That is, the code size is mostly assigned to the small probability events. In this case, it is applicable to use a smaller part of storage to save much more important information. ii) For information transmission: Compared with traditional communication, the transmission for big data has its own characteristics such as larger volume of data, a wide variety of events, and the value of information. Thus, it is sensible to preserve more information importance while reducing redundant information. In fact, the NMIM can be used as an efficient information importance measure to design rules for communication systems. iii) For information preprocessing: As for the information preprocessing, it is possible to analyze the performance of different divergences on distinguishing distinct distributions. Particularly, the MI divergence is a superior divergence in discerning a typical distribution from its adjacent distributions caused by rare events.

IV. APPLICATIONS IN DATA ANALYTICS OF IoTs

In the view of rare events analytics of IoTs, it is required to reduce the dimension as well as estimate the distributions and their functionals efficiently. That is, we should take methods to save more computing resources and improve the efficiency of data utilization [101]–[103]. Moreover, it is also necessary to analyze the relationships among rare events so that we can dig out more valuable information [104]–[106]. From the perspective of information theory, some approaches are discussed to deal with numerous information sources and do some data mining. Considering the relationship between information theory and big data analytics, we design an architecture based on information measures for rare events as shown in Fig. 5 whose details are summarized as follows:

i) Focusing on rare events: Rare events with small probability may contain more valuable information in some applications such as outlier detection and emergency alarm. In this case, it is necessary to define the rare events in a specific scenario at the first step.

ii) Selecting an information measure: An appropriate information measure can be adopted to characterize the distribution and highlight the importance of rare events. This is a mathematical representation of small probability events in the sense of the message importance.

iii) Dimension reduction and efficient estimation: As for the sample processing, it is essential to extract the most significant information with low dimension from the original data with high dimension. Especially, in the case of rare events, we can use low dimension samples and estimate the selected information measure to decrease the computation complexity.

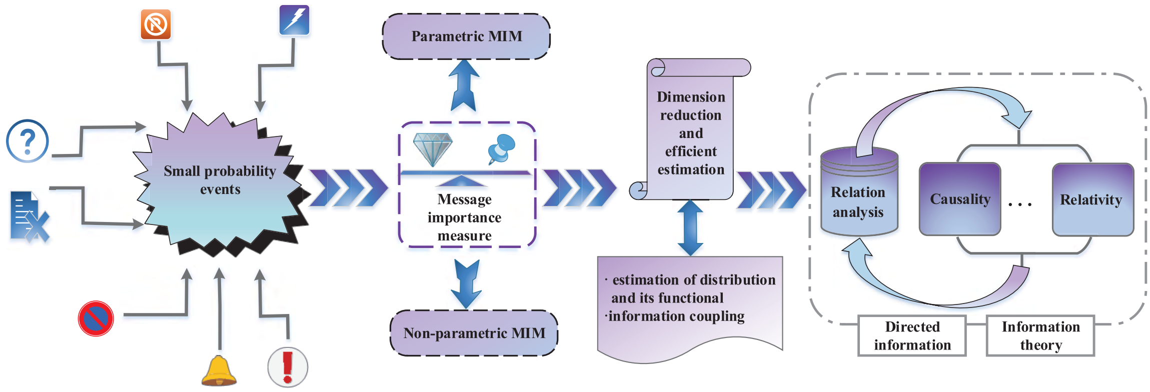


FIGURE 5. Architecture of data analytics based on message importance of rare events.

TABLE 3. Summary of literature on the efficient estimations of distribution and its functional.

Work Area	Related Work	Key Points
Estimation of Distribution	[111], [112], [113], [114], [115]	• Minimax estimation under ℓ_1 loss • Maximum likelihood estimator (MLE) based on risk functions • Non-asymptotic and asymptotic upper and lower bounds • Alphabet sizes and the number of samples both increasing
Estimation of Functionals of Distribution	[116], [117], [118], [119], [120], [121]	• Minimax estimator with the best polynomial approximation • Support size comparable with the number of samples • Non-smooth and smooth regimes for functionals
Entropy Estimation	[122], [123], [124], [125]	• Adaptive estimation framework • Dirichlet prior smoothing • Ensemble of plug-in estimators with weights
Information Divergence Estimation	[126], [127], [128], [129], [130], [131], [132], [133]	• Augmented plug-in estimator • Methodology with the polynomial approximation and the plug-in rule • Optimally weighted ensemble estimation

iv) Analyzing relationships: As for big data processing, it may be efficient to analyze the relationships among rare events by use of information measures.

In the architecture of data analytics for rare events, the information measures are discussed in the Section II. We shall specifically introduce some applications about how to use information measures in big data analytics as follows.

A. EFFICIENT ESTIMATION OF INFORMATION MEASURES

From the perspective of big data, it is quite essential to have efficient methods to estimate information measures, especially in the case of considerably large alphabet sizes. Whereas, the conventional estimation approaches can not work well [107]–[110], since that the rare events can not be observed accurately when the sample number is not very large. It is also worth investigating asymptotics with high dimension, especially when the number of samples is not much larger than the dimension. As a result, here lists some related works in the Table 3 whose details are described as follows.

- *Estimation of Distributions*: Based on some risk functions, different distribution estimations are investigated which play crucial roles in the information measure estimation [111]–[115]. For example, in the case that the alphabet size S increases with the number of samples n, a minimax estimation is discussed under the ℓ_1 loss (which is defined by $\mathbb{E}_P \sum_{i=1}^{S} |p_i - \hat{p}_i|$). This estimator has better performance on non-asymptotic upper and lower bounds of risk than maximum likelihood estimator (MLE).
- *Estimation of functionals of distribution*: When the unknown support size S is not smaller or even larger than the samples number n, a general methodology based on the minimax estimator is presented to estimate the functionals of distribution [116], [117]. Compared with the minimax estimator with non-smooth and smooth regions, the MLE is exactly sub-optimal in the large support [118]–[121].
- *Entropy Estimation*: As a widely used information measure, entropy is worth estimating especially. An adaptive

TABLE 4. Performance of minimax estimator and MLE and the comparison [117], [121].

Functional of distribution	Minimax squared error rates	Maximum squared error rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2}+\frac{\ln^2 n}{n}, (\frac{S}{\ln S}\lesssim n)$ [111], [117], [127], [128]	$\frac{S^2}{n^2}+\frac{\ln^2 n}{n}$ [121]
$H_\xi(P), 0<\xi\le\frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\xi}}, (\frac{S^{\frac{1}{\xi}}}{\ln S}\lesssim n, \ln n\lesssim \ln S)$ [111], [117]	$\frac{S^2}{n^{2\xi}}, (S^{\frac{1}{\xi}}\lesssim n)$ [121]
$H_\xi(P), \frac{1}{2}<\xi<1$	$\frac{S^2}{(n\ln n)^{2\xi}}+\frac{S^{2-2\xi}}{n}, (\frac{S^{\frac{1}{\xi}}}{\ln S}\lesssim n)$ [111], [117]	$\frac{S^2}{n^{2\xi}}+\frac{S^{2-2\xi}}{n}, (S^{\frac{1}{\xi}}\lesssim n)$ [121]
$H_\xi(P), 1<\xi<\frac{3}{2}$	$\frac{1}{(n\ln n)^{2-2\xi}}, (n\ln n\lesssim S)$ [111], [117]	$\frac{1}{n^{2-2\xi}}, (n\lesssim S)$ [121]
$H_\xi(P), \frac{3}{2}\le\xi$	$\frac{1}{n}$ [111], [117]	$\frac{1}{n}$ [111], [117], [121]

estimation framework is adopted to achieve the minimax rates in spite of the unknown support size S of distribution [122]. Besides, the estimator based on the best polynomial approximation also has the same performance [123]. Moreover, an inferior estimator is constructed by use of Dirichlet prior smoothing, which is similar to MLE but not as good as the above two [124]. In addition, an ensemble of plug-in estimators with weights is proposed to protect the results of estimation from decaying with the increase of sample dimension [125].

- *Information Divergences Estimation*: As a class of information measures, information divergences such as KL divergence, Hellinger distance and ℓ_2-divergence can be estimated in some similar ways [126]–[131]. In this regard, an augmented plug-in estimator and a methodology with the combination of polynomial approximation and plug-in rule are constructed to achieve the consistent estimator and the minimax rate-optimal estimator respectively [132]. Moreover, an optimally weighted ensemble estimator is also designed, which has good performance in the cases of high dimension [133].

In fact, the above classifications are based on the work areas of estimation. While, there exist some common criterions which can unify these estimators [117], [121], whose details are discussed as follows.

i) The maximum risk: Essentially, the MLE of distributions or their functionals complies with the maximum risk criterion which is given by

$$\sup_{P\in M_S}\mathbb{E}\{D_{error}(F(P)-\hat{F})\},\tag{4}$$

where D_{error} denotes a kind of error metric such as the one-norm and two-norm, $F(P)$ is a function of the distribution P whose support is M_S and $\hat{F}$ is the estimation for $F(P)$. In general, the MLE of distributions can be regarded as the fundamental plug-in estimator which is given by

$$\hat{p}_i=\frac{X_i}{n},\quad X_i=\sum_{j=1}^{n}I_{\{Z_j=i\}},\ (1\le i\le S),\tag{5}$$

where Z_j ($j\in\{1,2,...,n\}$) denotes the sample value, n is the sample number and S is the support size. Furthermore, we can substitute $\hat{p}_i$ into the functionals including $F(P)=P$ (namely the distribution itself) to obtain the estimation for the functionals of distribution. Moreover, as another example of MLE, the Dirichlet prior smoothing estimator is similar to plug-in estimator in the case of maximum squared risk, which is given by

$$\hat{P}_D=\frac{n}{n+\sum_{i=1}^{S}\alpha_i}\hat{P}+\frac{\sum_{i=1}^{S}\alpha_i}{n+\sum_{i=1}^{S}\alpha_i}\frac{\boldsymbol{\alpha}}{n+\sum_{i=1}^{S}\alpha_i},\tag{6}$$

where S is the alphabet size, $\hat{P}$ is an empirical distribution, and $\boldsymbol{\alpha}=(\alpha_1,\alpha_2...,\alpha_S)$ denotes the parameter vector which is adjustable. Besides, the ensemble of plug-in estimators with weights also belongs to MLE, which is defined by

$$\hat{F}_e=\sum_{l\in\bar{l}}\lambda_l\hat{F}_l,\quad(\sum_{l\in\bar{l}}\lambda_l=1),\tag{7}$$

where $\hat{F}_l$ is the plug-in estimator or its function, $\bar{l}=\{l_1,l_2,...,l_L\}$ is a set of parameters and λ_l denotes the weight value. In this estimator, the weights can be adjusted by using different optimal rules flexibly.

ii) The minimax risk: In terms of the minimax estimator for distributions or information functionals, it is based on the criterion minimizing the maximum risk of MLE which is given by

$$\inf_{\hat{F}}\sup_{P\in M_S}\mathbb{E}\{D_{error}(F(P)-\hat{F})\},\tag{8}$$

in which the notations are the same as those in the Eq. (4). As an instance of the minimax estimator, an approach based on the polynomial approximation rule is proposed, which treats the estimation problem as two cases of "small p_i" and "large p_i" (p_i denotes the probability element). In the case of "small p_i", the best polynomial approximation is used to guide the estimation, which is given by

$$P_K^*(x)=\arg\min_{P\in\Psi_K}\max_{x\in\Omega}|g(x)-P(x)|\tag{9}$$

where $g(x)$ is the objective function, Ψ_K is the set of polynomials with order no more than K on the domain Ω. Moreover, in the case of "large p_i", the estimation can be obtained by use of a kind of MLE such as the plug-in estimator.

Moreover, in order to see the reliability of the estimators based on these criterions (including the minimax risk or the maximum risk of MLE), it is necessary to compare the corresponding performance in some specific cases. Here, the results of estimating some classical information measures are summarized in the Table 4 in which

$H(P) = -\sum_{i=1}^{S} p_i \ln p_i$ denotes the Shannon entropy, $H_\xi(P) = \sum_{i=1}^{S} p_i^\xi$ ($\xi > 0$) is the dominant part of Renyi entropy, S is the support size, n is the samples' number, and the notation $a_k \lesssim b_k$ denotes $\sup_k \frac{a_k}{b_k} \le A$ (A is a constant). It is remarkable that the performance of the minimax estimator with n samples is equal to the MLE with $n \ln n$ samples in the case of small probability estimation, which is called ''effective sample size enlargement''.

B. DIMENSION REDUCTION BASED ON INFORMATION COUPLING

In the era of big data, there exists a big buzz word, ''dimension reduction'', which is involved in many fields such as machine learning, data mining, computer vision, etc. In order to solve this problem, more and more new techniques are being developed including principal component analysis, independent component analysis and regression analysis [134]–[137]. Besides, lots of applicable algorithms enable these new developed approaches to be used in many applications [138], [139]. However, these approaches are all designed from the viewpoint of the space of data rather than the intrinsic information flow.

On the contrary, the information coupling based on information measures is discussed to construct a framework for information-centric data processing. In fact, it is a novel view to analyze the information exchange process of relative data nodes by use of information coupling.

Mathematically, information coupling can be formulated in a fundamental communication scenario, where the input X contributes to the output Y through a transition probability matrix $W_{Y|X}$. In a typical communication system, a message U can form a Markov chain $U \to X \to Y$ with the input X and the output Y, where the message U is encoded into the input X. In order to design an efficient encoding scheme, it is usual to maximize the mutual information $I(U;Y)$ depending on the distribution P_U and the conditional distributions $P_{X|U=u}$. Similarly, the information coupling is to maximize the objective function $I(U;Y)$ constrained by a small mutual information $I(U;X)$. The constraints satisfy that the conditional distributions $P_{X|U}(\cdot|u)$ are neighbors of the marginal distribution P_X. That is, the information coupling [140] can be given by

$$\max_{U \to X \to Y} \frac{1}{n} I(U;Y), \tag{10}$$

$$\text{s.t. } \frac{1}{n} I(U;X) \le \sigma, \tag{10a}$$

$$\frac{1}{n} ||P_{X|U=u} - P_X||^2 = O(\sigma), \forall u, \tag{10b}$$

where the parameter σ is small enough.

In practice, the solution of the optimization problem about information coupling can provide a theoretical optimal result for dimension reduction from the perspective of information correlation. This can guide us to approximate the optimum by using low-dimensional information to represent the high-dimensional data. Specifically, suppose that there exists a hidden source sequence $\boldsymbol{x}_n = \{x_1, x_2, ..., x_n\}$ following the distribution P_X, an observed sequence $\boldsymbol{y}_n = \{y_1, y_2, ..., y_n\}$ following the distribution P_Y, and a transfer matrix $W_{Y|X}$ between the input X and output Y. In order to infer the hidden source X from Y, we usually require a sufficient statistic of $\boldsymbol{y}_n$ containing the whole information of $\boldsymbol{x}_n$. While, it is difficult to compute the statistic in the cases of the high dimensional structures of $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. To reduce the dimension, we would like to acquire a statistic from the observation $\boldsymbol{y}_n$ to characterize a certain feature of $\boldsymbol{x}_n$. According to the information coupling, a feature U in $\boldsymbol{x}_n$ is the most efficiently extracted from the observed data $\boldsymbol{y}_n$ in terms of the maximized mutual information $I(U;Y)$, which corresponds to the solution of this optimization problem. This efficient statistic based on the feature U, can be considered as a low-dimensional label containing the most significant information of the high-dimensional data, which implies an information theoretic method to reduce dimension [141].

Remark 2: Actually, it is not difficult to see that the information coupling is an efficient tool for statistics, which can extract the significant information from high dimensional original data. This can correspond to the goal of the dimension reduction and feature extraction for the rare events, which may use $\phi(U;X) = \mathcal{L}(\boldsymbol{p}_U) - \mathcal{L}(\boldsymbol{p}_X)$ to replace $I(U;X)$ to take the message importance transfer quantifying.

C. DIRECTED INFORMATION FOR RELATIONSHIP ANALYSIS

Directed information derived from information theory seems to be a commonly used approach, which can identify the interplay and causality between two stochastic processes [142]–[147]. Furthermore, it is also rational to adopt this approach to analyze the stochastic processes with rare events. Some details of directed information are given as follows.

In order to solve the causality problem in information systems [148], [149], an information measure, referred to as ''directed information'', is defined as

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}), \tag{11}$$

where $X^n = (X_1, X_2, \ldots, X_n)$ and $Y^n = (Y_1, Y_2, \ldots, Y_n)$ are independently random sequences, while X_i and Y_i ($i = 1, 2, ..., n$) are random variables, and $I(\cdot)$ denotes the mutual information. Moreover, due to the fact that the upper bound of the feedback channel capacity can be obtained by maximizing the normalized directed information [150], [151], another formulation of directed information is given by

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X_i; Y_i^n | X^{i-1}, Y^{i-1}). \tag{12}$$

which is obtained by use of the slide information (X^{i-1}, Y^{i-1}) [152].

Furthermore, this information measure has been adopted in some applications of relationship analysis, such as the computational biology with intrinsic causality [153], [154], the prediction of rate distortion [155] and the data compression with causal side information. Besides, the directed information provides an upper bound for the growth rates of optimal portfolios, which can also tightly bound the horse race gambling [142]. Notably, directed information can also measure the best error exponent for hypothesis testing which may be involved with the rare events identification.

Remark 3: Directed information is an efficient information measure which can interpret the causality transfer between two variables. Actually, this measure provides a significant tool to analyze the causal side information. Besides, it also plays an crucial role in dealing with the inference problem involved with causal influence factors. Similar method for the extension of MIM is necessary, which may bring some new insights on the massage importance discussion.

D. RARE EVENTS DETECTION FOR PROBABILITY DERIVATION PROCESS

In the data mining of IoTs, some scenarios such as urban abnormal pattern recognition as well as fire early warning and detection, can be treated as probability derivation processes which may be characterized and analyzed by means of information theory. It is worth noting that rare events detection lies in the intersection of the probability derivation process and the practical applications related to information measures. This problem has been investigated from many perspectives. In particular, the common methods of rare events detection are proposed based on the specific models or frameworks [156]–[160], such as Bayesian network anomaly detection, anomaly pattern classification in images, as well as normal behaviors definition for data points or groups.

As a typical probability derivation process, urban abnormal events detection is investigated widely, which may provide advices for governments and communities in smart city planning and management. In this regard, spatio-temporal data or multiple data sources are used to detect rare events of urban traffic states, such as mining uncommon trajectory of people, detecting road traffic anomalies [161], as well as identifying anomalous regions or locations [162]–[164]. The essential idea of these approaches is to construct a conditional probability model based on Hidden Markov process or Maximum Likelihood rule to detect or predict anomalous events. That is, the underlying distribution of rare patterns can be obtained in the probabilistic models which are constructed based on the different patterns of spatio-temporal data.

Moreover, message measures based on similarity and correlation also play crucial roles in identifying urban abnormal events [165], [166]. For instance, $L_{-\infty}$ distance is adopted as a kind of similarity measurement to evaluate the degree of anomalous traffic [167]. Besides, KL divergence is also commonly used as a metric to measure correlation [168], [169]. In video surveillance systems of urban traffic states, when a small video clip is represented as a histogram of multi-set bag of codewords by using Fourier based trajectory feature descriptor [168], KL divergence is applied to classify the pending video clips into the normal or abnormal ones. The corresponding metric based on KL divergence is given by

$$D_{KL}(Q||P_0) - D_{KL}(Q||P_1) = \ln \prod_{i=1}^{K-1} \left(\frac{p(v_i|c=1)}{p(v_i|c=0)}\right)^{q_i}, \quad (13)$$

where $D_{KL}(\cdot)$ denotes the operator of KL divergence, $p(v_i|c=1)$ and $p(v_i|c=0)$ are probability elements from the codewords of normal video clips and abnormal ones (the corresponding distributions are P_1 and P_0), q_i denotes the probability element from the codewords of pending video clips (the corresponding distribution is Q). Furthermore, a spatio-temporal detector for the mixture of dynamic textures (MDT) model is proposed, in which the center-surround saliency detection is based on the KL divergence between feature responses and events class labels [169]:

$$D_{KL}(P_{X|c}||P_X) \doteq \sum_i \left\{\pi_i^c \log \frac{\sum_j^{K_c} \pi_j^c e^{(-D_{KL}(p^i_{X|c}||p^j_{X|c}))}}{\sum_j^{K_0+K_1} \omega_j e^{(-D_{KL}(p^i_{X|c}||p^j_{X}))}}\right\}, \quad (14)$$

where $p^i_{X|c}$ are class-conditional densities (based on the class $c \in \{0, 1\}$), p^j_X are sample densities, π_j^c and ω_j are parameters, K_c ($c \in \{0, 1\}$) denotes the number of samples in the corresponding class c.

Similar to the KL divergence, the message measures mentioned in Section II may be also efficient in rare events detection for spatio-temporal data and may perform better in some special data sets, which can be investigated further in probability derivation processes.

Remark 4: Some message measures reveal the similarity or correlation for probability derivation processes. Specifically, these measures can be regarded as criteria for urban abnormal events mining. In general, it is promising to make good use of novel information measures to extend the strategies of rare events detection.

V. FUTURE CHALLENGES

Considering future research directions, new approaches and challenging applications can promote the development of information measures with respect to rare events. By combining big data analytics, an architecture of rare events processing based on information measures is constructed shown in Fig. 6. In particular, we can apply big data analytics and information measures in the challenging scenarios involved with rare events, including smart cities, autonomous driving, and detection in IoTs.

Actually, in the above applications, the common technique playing a core role is rare events detection. Here, we design a technology framework in the viewpoint of information measures to help to detect rare events as shown in Fig. 7. To be specific, assume there exist two different kinds of message sequences in the data set, that is, the data set consists of two message sources X and Y with different distributions. In this

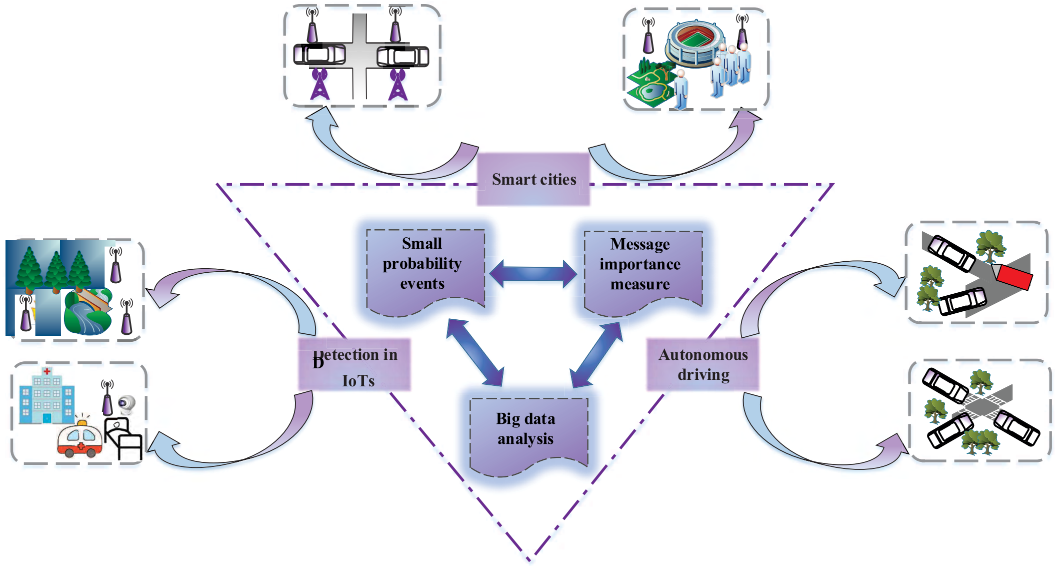


FIGURE 6. Architecture of small probability events (namely rare events) processing based on information measures for challenging applications.

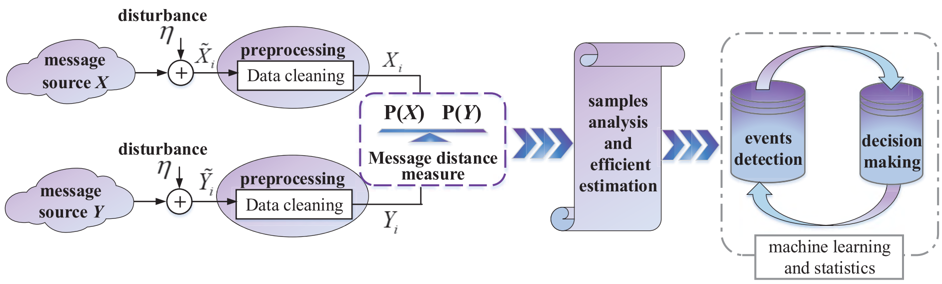


FIGURE 7. The framework for rare events detection.

case, the message sequences from the message source Y are considered as the rare events. The goal of our framework is to detect message sequences of Y. Our core idea is to make use of information measures such as KL divergence, Renyi divergence and f-divergence to identify the two kinds of information distributions. In this case, we assume that how to design efficient information measures is a fundamental problem in the first step. Moreover, when an information measure is obtained, we also need to analyze the samples in the message sequences and take efficient methods to estimate the information measure. Furthermore, it is applicable to classify estimated results by resorting to the machine learning algorithms so that we can make a decision for rare events detection.

In addition, it is promising to measure rare events based on message importance and then analyze the relationship among the big data. The emerging applications related to big data require new ways to deal with anomalous detection or probability events mining. To this end, we summarize some challenges and perspectives associated with rare events processing, which can be future research directions for information measures as shown in Table 5.

A. SMART CITIES

1) ANOMALY DETECTION FOR URBAN MONITORING DATA

As a typical application of big data, smart city has been evolving rapidly with the increase of urban population. This implies that cities can be monitored by countless devices in many aspects such as road traffic, transportation management, environment monitoring, healthcare, etc. Actually, in cities, it is significant to detect the anomalies with small probability, which may provide effective guidance or warning information.

In order to investigate the anomaly detection problem in smart cities [170], the major challenges are listed as follows.

- Security problems in the urban monitoring systems with wireless sensor networks (WSN).
- The way to optimize the validity and reliability of transportation schedule system by avoiding the anomalies.
- The long time prediction for the regular pattern of cities.
- To distinguish the unexpected events from popular anomalies.
- Automatic anomaly detection algorithms for the urban monitoring systems with IoTs.

TABLE 5. Perspective applications and use cases of rare events processing.

Work Area	Related Work	Key Points
Smart cities	[41], [170], [171], [172], [173], [174], [175], [176], [177], [178], [179], [180], [181], [182], [183], [184], [185], [186], [187],	• Anomaly detection with urban surveillance data: ◦ Security problems in wireless sensor networks ◦ Optimizing transportation schedule system ◦ The prediction problem in cities • Urban Black Holes detection: ◦ Detecting the groups of objects described by graphs ◦ Multiple datasets analysis
Automatic driving	[188], [189], [190], [191], [192], [193], [194], [195], [196], [197], [198], [199], [200], [201], [202], [203], [204], [205], [206], [207], [208],	• Obstacles detection • Detecting and tracking objects • Detecting road surfaces and lanes • probabilistic approaches and learning strategies
Applications to detection in IoTs	[209] [210], [211], [212], [213], [214], [215], [216], [217], [218], [219], [220], [221], [222]	• Differentiating normal and abnormal data by use of judging criterion • Machine learning algorithms such as classification, clustering and R-PCA

In fact, the anomaly detection can be processed in many ways including machine learning, signal analysis and even information theory. To be honest, there are some specific methods to detect the anomalies in smart cities, which may overcome the above challenges from different perspectives. Particularly, in order to improve the security of the WSNs in urban monitoring systems, a non-intrusive architecture is proposed to detect attacks by use of the support vector machine (SVM) [171]. Moreover, for the IoTs of smart cities, by using automatic clustering or classification, the events with low probability can be identified in many applications such as the car parking scenario, polluted region monitoring and actionable bumps detection [41]. In a wide sense, automatic target detection in urban surveillance systems can be also regarded as an anomaly detection [172]. As an example, there is a method presented to extract the regions with the highest energy frequency in pending images, which can help to reduce the complexity of detection. In addition, spatio-temporal data mining is also considered in urban anomaly detection. Specifically, a two-step method (to compute individual anomaly scores (CIAS) and to aggregate the individual anomaly scores (AIAS)) is proposed to give an anomaly score for each data source of each region at each time slot [173]; An improved Local Outlier Factor (LOF) algorithm (based on spatial-temporal cube) is adopted for abnormal region detection [174]; A Urban Anomaly PreDiction (UAPD) framework is designed to detect the anomalous change points and dig out the time-evolving inherent factors [175].

Notably, it is promising to exploit information theory to deal with anomaly detection by emphasizing the importance of rare events. By combining machine learning techniques, the importance measures focusing on rare events may provide new ways to cope with the anomaly detection and the evaluation of post processing, which plays an vital role in smart cities.

2) DETECTING URBAN BLACK HOLES

As an important part of smart cities, the urban black hole denotes a region in which the whole traffic inflow is larger than the whole traffic outflow. Actually, the urban black hole can reflect emergencies or irregular events, namely rare events, including disasters, accidents, as well as traffic jams or congestion [176], [177]. It is worth detecting urban black holes efficiently, which can make a beneficial effect on urban safety. Therefore, some approaches are investigated as follows.

- Graph Clustering: With regard to the graph clustering, the approaches with the pruning schemes and the random matrix are proposed to characterize the potential black holes in a directed graph [178]. Besides, there are some other approaches detecting black holes by means of different measures [179], [180] such as attribute, modularity and density.
- Dynamic Graph Detection: To detect black holes emerging in dynamic graph, some efficient approaches are proposed by means of the increment, pattern trees, and the pattern recognition with constraints [181], [182].
- Groups Moving Recognition: On one hand, the density of regions is used to discover the object groups beyond the threshold during the observation time [183]. On the other hand, moving together behaviors during a given time period are investigated to find out the tracking of a group of objects [184], [185].

- Spatio-Temporal Graph: Based on the spatio-temporal graph, some approaches are presented to mine spatial urban black holes [186], as well as, detect the tracking of data temporally and spatially [187].

Actually, from the perspective of probability distribution, it is possible to use information measures to find out urban black holes which may be described by graph methods. To do so, a detection scheme for the smart city is shown in Fig. 8 whose details are as follows.

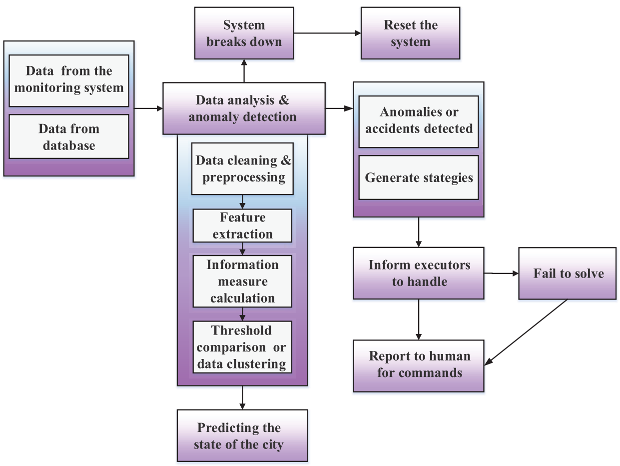


FIGURE 8. The rare events detection scheme for the smart city.

Specifically, data from the monitoring system and database are used to detect the emergency events or accidents which can be regarded as the rare events. Next, we can apply the information measures to analyze the relationships among events. By using the data analysis and processing, the control center predicts the state of the city. This can be adopted as a reference to update the system model and database. Moreover, when anomalies are detected or the system breaks down, the control center can reset the system. If anomalies or accidents are detected or some unsolved emergencies are reported, control center will take measures to handle them. Then, if the system generates strategies for the anomaly detection, it would send commands to executors to solve the problems. Besides, human can also set in the work directly when the system fails to finish the work.

B. AUTONOMOUS DRIVING

As an important part of the autonomous driving, obstacles detection makes a great influence on warning and predicting collisions and accidents [188]. However, it is still a challenge to accurately detect the obstacles or objects with small probability in the view of computer vision. In general, some key issues of autonomous driving are summarized as follows.

- Obstacles Detection: On one hand, some approaches are presented to characterize obstacles by use of image data [189], v-disparity histogram [190], as well as the models for the height-over ground [191]–[193]. On the other hand, deep learning tasks are used to detect obstacles by means of the image features and related information. Moreover, a technique ‘‘6D-vision’’ is also put forward to discover the dangerous events on the roads [194], [195].
- Object Detection: There are some approaches to detect and track objects by means of classification or clustering [196]. The strategies and frameworks for object localization or tracking are also proposed depending on the Kalman filter [197] and deep convolutional neural networks [198]. Furthermore, some other approaches are designed by use of the trade-off between camera orientations prediction and monitoring techniques [199], [200].
- Detecting Road Surface and Lanes: As for road surface detection, the discriminant analysis (DA) is presented to characterize the road crack [201], [202]. This can provide a threshold for classification according to the road texture and color in images. Besides, in order to detect the road curb and lanes, it is common to regard color and texture as interesting features of roads. These can be used by combining classification with the hue-saturation-intensity (HSI) color space or red-green-blue (RGB) color space [203], [204]. Besides, another framework of road curb and lanes detection is addressed by extracting the 3D parameters from some curb models [205].

Moreover, there are some works proposed based on probabilistic approaches and learning strategies. Gaussian process (GP) regression decomposition based on a superpixel-like algorithm is employed to validate quasi-constant velocity models which build a set of Kalman filters to identify the abnormal motions online [206]. A particle swarm optimization (PSO) and bacterial foraging optimization (BFO)-based learning strategy (PBLS) is presented to improve the classifier and loss function of strengthened region proposal network (SRPN), which can be applied in object detection of autonomous driving [207]. A set of 3D object proposals based on an energy function are obtained to detect high-quality 3D objects by use of a convolutional neural net (CNN) [208].

Additionally, with regard to the detection for autonomous driving, it is apparent that rare events play important roles in many aspects of vehicular safety system. By measuring small probability, it is appropriate to apply information theory to the autonomous driving detection. To do so, an obstacle detection scheme is shown in Fig. 9, whose details are given as follows. Based on the data from monitoring devices or radars, the autonomous control system can detect obstacles or other outlier events, which can be analyzed by use of information measures. If no obstacle is discovered, the system will continue the normal surveillance. However, if some obstacles are detected, the system will take measures to solve the problem by slowing down and choosing a new way. When the emergencies are not solved well, it will put on the brake and report them to drivers for further commands.

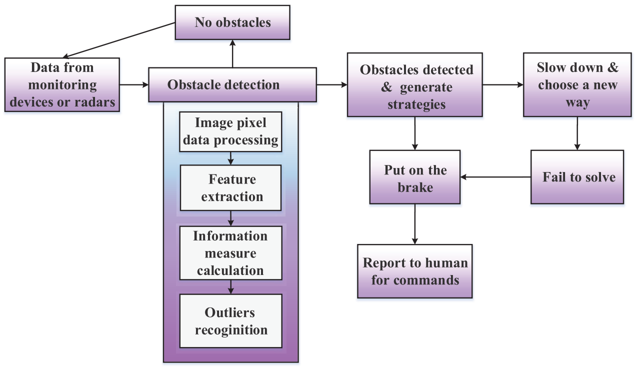


FIGURE 9. **The obstacle detection scheme for the autonomous driving.**

C. APPLICATIONS ON DETECTION IN IOTS

Outlier detection in IoTs, is to dig out the minority of sensors data exactly [209], [210]. In fact, it is essential to differentiate the outlier data or observations from the normal data so that one can gain the warning information and prevent the outlier data from misleading us [211]. There exist a various of researches focusing on the outlier detection which are also considered to detect rare events in IoTs systems.

On one hand, there are some approaches to detect outliers in IoTs directly, such as using Jaccard coefficient or Euclidean distance as the criterion of decision making [212], referring to the expert knowledge on security [213], as well as, monitoring the abnormal traffic among communication devices [214], etc. On the other hand, several researches divide observations into different groups to find out the outliers by use of classification and clustering algorithms [215], [216]. To address this kind of matter, a few approaches also introduce static data series [217] or dynamic time series into the machine learning algorithms. Besides, a framework of data analysis is put forward by means of the recursive principal component analysis (R-PCA) [209], which provides another way to investigate the security of IoTs systems.

In light of the fact that the data observed from IoTs are usually fed to cloud service systems, some approaches are proposed by blending both IoTs and cloud technologies [218]. Moreover, to test IoTs systems conveniently, a new method is presented to emulate the environments of IoTs by means of a network emulator, which can improve the processing efficiency for outlier detection [219].

Furthermore, some probabilistic models and large-scale processing approaches are also exploited in the anomalies detection of IoTs. A statistical decision framework based on temporally correlated traffic is designed, which develops two low-complexity algorithms (based on cross entropy method and generalized likelihood ratio test) to achieve anomaly detection and attribution [220]. An adversarial statistical learning mechanism, outlier Dirichlet mixture-based anomaly detection systems (ODM-ADS), is presented to obtain legitimate profiles and discover suspicious anomalies [221]. Besides, there are two methods are proposed, namely a one-class support Tucker machine (OCSTuM) and an OCSTuM based on a genetic algorithm called GA-OCSTuM, which extend one-class support vector machines to tensor space to detect anomalies in IoTs [222].

However, in spite of many efficient approaches for outlier detection, few researches consider to exploit the small probability character in the viewpoint of probability distribution. Actually, it is promising to take use of information measures to analyze the outliers of IoTs.

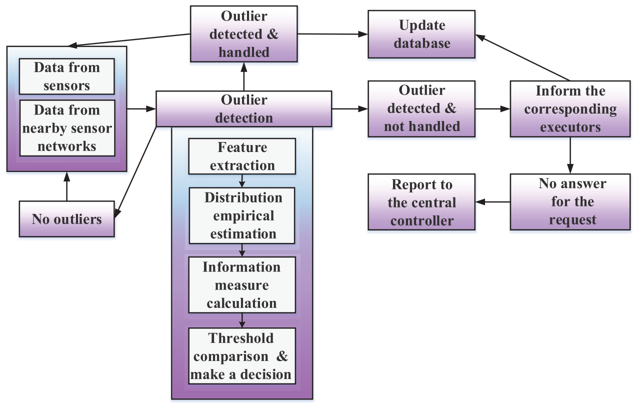


FIGURE 10. **The outlier detection scheme for the IoTs composed of monitoring sensors.**

From the perspective of information theory, importance measures can provide a specific access to tackle the outlier detection problem by using probability distribution, which is shown in Fig. 10 whose details are as follows. The data collected by distributed sensors are used to detect the potential or ongoing outliers by resorting to information measures. If an outlier is detected and handled, the local center will continue to collect data and update the database. However, if a detected outlier is not handled well, the local center will contact with executors to solve the problem and save the data to the database. Once there is no answer for the request, local center will report it to the control center.

VI. CONCLUSION

In this paper, we gave a total review on information measures for rare events in big data. In order to characterize the importance of rare events, we summarized some message measures such as the parametric MIM and the non-parametric MIM which have properties on emphasizing small probability elements for a given distribution. These information measures are regarded as promising criterions or tools for statistical big data analytics. Furthermore, we introduced that measures focusing on rare events can provide new ways for message processing such as compression and transmission. Moreover, some other applications in big data have been discussed including efficient estimation, dimension reduction and relationship analysis. Additionally, we introduced that information measures for rare events could be applicable for some future research directions including smart cities, autonomous driving, and anomaly detection in IoTs. In these cases, there exist several future challenges of information measures summarized as follows:

i) Data storage and low latency computation for the data sets containing rare events.

ii) Feature extraction and data cleaning of holding rare events.

iii) Design of information theoretic criterions to measure distributions while considering the values of rare events.

iv) Efficient methods of information measure estimations.

v) Correlation and causality analysis based on information measures.

vi) Decision making strategies for rare events (or probability events) mining.

REFERENCES

[1] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT data analytics: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.

[2] X.-W. Chen and X. Lin "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[3] H. Hu, Y. Wen, T. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 5, pp. 7776–7797, Jun. 2017.

[4] A. Leureux, K. Grolinger, H. Elyamany, and M. Capertz, "Machine learning with big data: Challenges and Approahces," *IEEE Access*, vol. 5, pp. 2169–3536, Apr. 2017.

[5] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, Oct. 2014.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[7] H. Yu, Z. H. Tan, Y. Zhang, Z. Ma, and J. Guo, "DNN filter bank cepstral coefficients for spoofing detection," *IEEE Access*, vol. 5, pp. 4779–4787, 2017.

[8] Z. Ma, H. Yu, Z.-H. Tan, and J. Guo, "Text-independent speaker identification using the histogram transform model," *IEEE Access*, vol. 4, pp. 9733–9739, 2017.

[9] Y. Lu, K. Xiong, P. Fan, Z. Zhong, and K. Letaief, "Robust transmit beamforming with artificial redundant signals for secure SWIPT system under non-linear EH model," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2218–2232, Jan. 2018.

[10] K. Xiong, C. Chen, G. Qu, P. Fan, and K. B. Letaief, "Group cooperation with optimal resource allocation in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3840–3853, Jun. 2017.

[11] K. Xiong, P. Fan, Y. Lu, and K. Letaief, "Energy efficiency with proportional rate fairness in multi-relay OFDM networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1431–1447, May 2016.

[12] Y. Lu, K. Xiong, P. Fan, Z. Ding, Z. Zhong, and K. Letaief, "Global energy efficiency in secure MISO SWIPT systems with non-linear power-splitting EH model," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 216–232, Jan. 2019.

[13] B. Koo, S. Lee, M. Lee, D. Lee, S. Lee, and S. Kim, "PDR/fingerprinting fusion indoor location tracking using RSS recovery and clustering," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Busan, South Korea, Sep. 2015, pp. 699–704.

[14] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[15] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

[16] Z. Ma, A. Leijon, and W. B. Kleijn, "Vector quantization of LSF parameters with a mixture of Dirichlet distributions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1777–1790, Sep. 2013.

[17] H. Gao, X. Wang, J. Tang, and H. Liu, "Network denoising in social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Niagara Falls, ON, Canada, Apr. 2014, pp. 564–571.

[18] Y. Qi, Y. Song, and T. Xiang, "Making better use of edges via perceptual grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Oct. 2015, pp. 1856–1865.

[19] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[20] F. Pourkamali-Anaraki and S. Becker, "Preconditioned data sparsification for big data with applications to PCA and K-means," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2954–2974, May 2017.

[21] S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *Proc. IEEE Intell. Comput. Technol. Automat. (ICICTA)*, Madurai, India, May 2010, pp. 50–53.

[22] M. S. Beasley, J. V. Carcello, D. R. Hermanson, and P. D. Lapides, "Fraudulent financial reporting: Consideration of industry traits and corporate governance mechanisms," *Accounting Horizons*, vol. 14, no. 4, pp. 441–454, Dec. 2000.

[23] W. Lee and S. Stolfo, "Data Mining Approaches for Intrusion Detection," in *Proc. Usenix Secur.*, San Antonio, TX, USA, Jan. 1998, pp. 291–300.

[24] K. Julisch and M. Dacier, "Mining intrusion detection alarms for actionable knowledge," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2002, pp. 366–375.

[25] X. Zhang and X. Hao, "Research on intrusion detection based on improved combination of K-means and multi-level SVM," in *Proc. IEEE 17th Int. Conf. Commun. Technol. (ICCT)*, Chengdu, China, Oct. 2017, pp. 2042–2045.

[26] M. Li, "Application of CART decision tree combined with PCA algorithm in intrusion detection," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Beijing, China, Nov. 2017, pp. 38–41.

[27] G. Karatas and O. Sahingoz, "Neural network based intrusion detection systems with different training functions," in *Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS)*, Antalya, Turkey, Mar. 2018, pp. 1–6.

[28] V. Desnitsky and I. Kotenko, "Event analysis for security incident management on a perimeter access control system," in *Proc. 19th IEEE Int. Conf. Soft Comput. Meas. (SCM)*, St. Petersburg, Russia, May 2016, pp. 481–483.

[29] D. D. Hwang, P. Schaumont, K. Tiri, and I. Verbauwhede, "Securing embedded systems," *IEEE Security Privacy*, vol. 4, no. 2, pp. 40–49, Mar. 2006.

[30] S. Ravi, A. Kocher, and S. Hattangady, "Security in embedded systems: Design challenges," *ACM Trans. Embedded Comput. Syst.*, vol. 3, no. 3, pp. 461–491, Aug. 2014.

[31] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the Internet of Things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[32] J. Wu and W. Zhao, "Design and realization of winternet: From net of things to Internet of Things," *ACM Trans. Cyber Phys. Syst.*, vol. 1, no. 1, Feb. 2017, Art. no. 2.

[33] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[34] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of Things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, Mar. 2016.

[35] H. Sedjelmaci, S. M. Senouci, and M. Al-Bahri, "Alightweight anomaly detection technique for low-resource IoT devices: A game-theoretic methodology," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[36] P. Souza, W. Marques, F. Rossi, G. Rodrigues, and R. Calheiros, "Performance and accuracy trade-off analysis of techniques for anomaly detection in IoT sensors," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Da Nang, Vietnam, Jan. 2017, pp. 486–491.

[37] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time intrusion detection in the Internet of Things," *Ad Hoc Netw.*, vol. 11, no. 8, pp. 2661–2674, May 2013.

[38] H. Sedjelmaci, S. M. Senouci, and M. Al-Bahri, "A lightweight anomaly detection technique for low-resource IoT devices: A game-theoretic methodology," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[39] J. C. Bezdek, S. Rajasegarar, M. Moshtaghi, C. Leckie, M. Palaniswami, and T. C. Havens, "Anomaly detection in environmental monitoring networks," *IEEE Comput. Intell. Mag.*, vol. 6, no. 2, pp. 51–58, Apr. 2011.

[40] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[41] R. Jain and H. Shah, "An anomaly detection in smart cities modeled as wireless sensor network," in *Proc. Int. Conf. Signal Inf. Process. (IConSIP)*, Nanded, India, Oct. 2016, pp. 1–5.

[42] O. Kotevska, A. G. Kusne, D. V. Samarov, A. Lbath, and A. Battou, "Dynamic network model for smart city data-loss resilience case study: City-to-city network for crime analytics," *IEEE Access*, vol. 5, pp. 20524–20535, Oct. 2017.

[43] E. Markova, I. Gudkova, A. Ometov, I. Dzantiev, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Flexible spectrum management in a smart city within licensed shared access framework," *IEEE Access*, vol. 5, pp. 22252–22261, 2017.

[44] S. Mallapuram, N. Ngwum, F. Yuan, C. Lu, and W. Yu, "Smart city: The state of the art, datasets, and evaluation platforms," in *Proc. 16th IEEE/ACIS Int. Conf. Comput. Inf. Sci. (ICIS)*, Wuhan, China, May 2017, pp. 447–452.

[45] S. Wan, J. Lu, P. Fan, and K. B. Letaief, "To smart city: Public safety network design for emergency," *IEEE Access*, vol. 6, pp. 1451–1460, Dec. 2017.

[46] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, 2000.

[47] F. Harrou, F. Kadri, S. Chaabane, C. Tahon, and Y. Sun, "Improved principal component analysis for anomaly detection: Application to an emergency department," *Comput. Ind. Eng.*, vol. 88, pp. 63–77, Oct. 2015.

[48] S. Xu, M. Baldea, T. F. Edgar, W. Wojsznis, T. Blevins, and M. Nixon, "An improved methodology for outlier detection in dynamic datasets," *AIChE J.*, vol. 61, no. 2, pp. 419–433, 2015.

[49] H. Yu, F. Khan, and V. Garaniya, "Nonlinear Gaussian belief network based fault diagnosis for industrial processes," *J. Process Control*, vol. 35, pp. 178–200, Nov. 2015.

[50] A. Prieto-Moreno, O. Llanes-Santiago, and E. García-Moreno, "Principal components selection for dimensionality reduction using discriminant information applied to fault diagnosis," *J. Process Control*, vol. 33, pp. 14–24, Sep. 2015.

[51] US Department of Transportation, National Highway Traffic Safety Administration, Washington, DC, USA. *Traffic Safety Facts 2011*. Accessed: 2013. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811754

[52] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, "Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Los Angeles, CA, USA, Jun. 2017, pp. 1025–1032.

[53] P. Amaradi, N. Sriramoju, L. Dang, G. S. Tewolde, and J. Kwon, "Lane following and obstacle detection techniques in autonomous driving vehicles," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, Grand Forks, ND, USA, May 2016, pp. 0674–0679.

[54] V. Gaikwad and S. Lokhande, "An improved lane departure method for advanced driver assistance system," in *Proc. Int. Conf. Comput., Commun. Appl. (ICCCA)*, Dindigul, India, Feb. 2012, pp. 1–5.

[55] S. Lee, H. Son, and K. Min, "Implementation of lane detection system using optimized Hough transform circuit," in *Proc. IEEE Asia–Pacific Conf. Circuits Syst. (APCCAS)*, Kuala Lumpur, Malaysia, Dec. 2010, pp. 406–409.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[57] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[58] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[59] W. Song, Y. Yang, M. Fu, F. Qiu, and M. Wang, "Real-time obstacles detection and status classification for collision warning in a vehicle active safety system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 758–773, Mar. 2017.

[60] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[61] Q. Jiang, B. Wang, and X. Yan, "Multiblock independent component analysis integrated with Hellinger distance and Bayesian inference for non-Gaussian plant-wide process monitoring," *Ind. Eng. Chem. Res.*, vol. 54, no. 9, pp. 2497–2508, 2015.

[62] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2001, pp. 130–143.

[63] S. Ando and E. Suzuki, "An information theoretic approach to detection of minority subsets in database," in *Proc. IEEE 6th Int. Conf. Data Mining*, Hong Kong, China, Dec. 2006, pp. 11–20.

[64] A. Anderson and H. Haas, "Kullback-Leibler Divergence (KLD) based anomaly detection and monotonic sequence analysis," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.

[65] H. Touchette, "The large deviation approach to statistical mechanics," *Phys. Rep.*, vol. 478, nos. 1–3, pp. 1–69, Jul. 2009.

[66] R. P. Curiel and S. Bishop, "A measure of the concentration of rare events," *Sci. Rep.*, vol. 6, no. 32369, pp. 1–6, Aug. 2016.

[67] N. Weinberger and N. Merhav, "A large deviations approach to secure lossy compression," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2533–2559, Apr. 2017.

[68] I. C. Paschalidis and G. Smaragdakis, "A large deviations approach to statistical traffic anomaly detection," in *Proc. IEEE Conf. Decis. Control (CDC)*, San Diego, CA, USA, Dec. 2006, pp. 1900–1905.

[69] A. Youssef, C. Delpha, and D. Diallo, "Analytical model of the KL Divergence for Gamma distributed data: Application to fault estimation," in *Proc. IEEE Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 1–6.

[70] L. Li, Q. Xu, X. Luo, and S. Sun, "Key frame selection based on KL-divergence," in *Proc. IEEE Int. Conf. Multimedia Big Data (BigMM)*, Beijing, China, Apr. 2015, pp. 1–6.

[71] Z. Youbi, L. Boubchir, M. D. Bounneche, A. Ali-Chérif, A. Boukrouche, "Human Ear recognition based on multi-scale local binary pattern descriptor and KL divergence," in *Proc. IEEE Telecommun. Signal Process. (TSP)*, Vienna, Austria, Jun. 2016, pp. 1–6.

[72] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3884–3890, Jul. 2010.

[73] R. She, S. Y. Liu, and P. Fan, "Amplifying inter-message distance: On information divergence measures in big data," *IEEE Access*, vol. 5, pp. 24105–24119, 2017.

[74] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 988–992.

[75] J. Cao, Z. Wu, B. Mao, and Y. Zhang, "Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system," *World Wide Web*, vol. 16, nos. 5–6, pp. 729–748, 2013.

[76] G. Zhu, J. Cao, C. Li, and Z. Wu, "A recommendation engine for travel products based on topic sequential patterns," *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 17595–17612, 2017.

[77] J. Cao, Z. Wu, Y. Wang, and Y. Zhuang, "Hybrid collaborative filtering algorithm for bidirectional Web service recommendation," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 607–627, 2013.

[78] P. Fan, Y. Dong, J. X. Lu, and S. Y. Liu, "Message importance measure and its application to minority subset detection in big data," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[79] R. She, S. Liu, and P. Fan, "Recognizing information feature variation: Message importance transfer measure and its applications in big data," *Entropy*, vol. 20, no. 6, pp. 1–22, May 2018.

[80] S. Liu, Y. Dong, P. Fan, R. She, and S. Wan, "Matching users' preference under target revenue constraints data recommendation systems," *Entropy*, vol. 21, no. 2, p. 205, Feb. 2019.

[81] S. Liu, R. She, and P. Fan, "Differential message importance measure: A new approach to the required sampling number in big data structure characterization," *IEEE Access*, vol. 6, pp. 42851–42867, 2018.

[82] S. Wan, J. Lu, P. Fan, and K. B. Letaief, "Minor probability events,"detection in big data: An integrated approach with bayes detection and MIM," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 418–421, Mar. 2019.

[83] R. She, S. Y. Liu, Y. Q. Dong, and P. Fan, "Focusing on a probability element: Parameter selection of message importance measure in big data," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[84] S. Liu, R. She, P. Fan, and J. Lu, "Non-parametric message important measure: Compressed storage design for big data in wireless communication systems," in *Proc. IEEE Asia–Pacific Conf. Commun. (APCC)*, Perth, WA, Australia, Dec. 2017, pp. 1–6.

[85] S. Liu, R. She, P. Fan, and K. B. Letaief, "Non-parametric message importance measure: Storage code design and transmission planning for big data," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5181–5196, Nov. 2018.

[86] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*. Cham, Switzerland: Springer, 2014.

[87] M. Gharbieh, H. ElSawy, A. Bader, and M.-S. Alouini, "Spatiotemporal stochastic modeling of IoT enabled cellular networks: Scalability and stability analysis," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3585–3600, Aug. 2017.

[88] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, Oct. 2015.

[89] I. Estella-Aguerri and A. Zaidi, "Lossy compression for compute-and-forward in limited backhaul uplink multicell processing," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5227–5238, Dec. 2016.

[90] T. Cui, L. Chen, and T. Ho, "Distributed distortion optimization for correlated sources with network coding," *IEEE Trans. Commun.*, vol. 60, no. 5, pp. 1336–1344, May 2012.

[91] S. Jalali and T. Weissman, "Block and sliding-block lossy compression via MCMC," *IEEE Trans. Commun.*, vol. 60, no. 8, pp. 2187–2198, Aug. 2012.

[92] A. Sechelea, A. Munteanu, S. Cheng, and N. Deligiannis, "On the rate-distortion function for binary source coding with side information," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5203–5216, Dec. 2016.

[93] S. Verdú, "Fifty years of Shannon theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2057–2078, Oct. 1998.

[94] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing), 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[95] T. V. Erven and P. Harremoes, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jun. 2014.

[96] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. New York, NY, USA: Springer, 1998, pp. 199–213.

[97] B. Masnick and J. Wolf, "On linear unequal error protection codes," *IEEE Trans. Inf. Theory*, vol. 3, no. 4, pp. 600–607, Oct. 1967.

[98] D. Sejdinovic, D. Vukobratovic, A. Doufexi, V. Senk, and R. J. Piechocki, "Expanding window fountain codes for unequal error protection," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2510–3526, Sep. 2009.

[99] K. Sun and D. Wu, "Unequal error protection for video streaming using delay-aware fountain codes," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.

[100] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testin," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, Jul. 2014.

[101] K. Lo and H. Lin, "Cost-Sensitive encoding for label space dimension reduction algorithms on multi-label classification," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Taipei, Taiwan, Dec. 2017, pp. 136–141.

[102] W. Li, F. Feng, H. Li, and Q. Du, "Discriminant analysis-based dimension reduction for hyperspectral image classification: A survey of the most recent advances and an experimental comparison of different techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 15–34, Mar. 2018.

[103] J. Zhang, J. Yu, and D. Tao, "Local deep-feature alignment for unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2420–2432, May 2018.

[104] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, May 2016.

[105] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[106] X. Yang, P. Zhao, X. Zhang, J. Lin, and W. Yu, "Toward a Gaussian mixture model-based detection scheme against data integrity attacks in the smart grid," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 147–161, Feb. 2017.

[107] J. Hájek, "A characterization of limiting distributions of regular estimates," *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 14, no. 4, pp. 323–330, 1970.

[108] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, Berkeley, CA, USA, vol. 1, 1972, pp. 175–194.

[109] L. Le Cam, *Asymptotic Methods Statistical Decision Theory*. New York, NY, USA: Springer, 1986.

[110] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2200–2203, Sep. 2004.

[111] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 2291–2295.

[112] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6343–6354, Nov. 2015.

[113] D. L. Donoho and I. M. Johnstone, "Minimax risk over l_p-balls for l_q-error," *Probab. Theory Rel. Fields*, vol. 99, no. 2, pp. 277–303, 1994.

[114] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[115] Diakonikolas, "Beyond histograms: Structure and distribution estimation," in *Proc. STOC Workshop*, New York, NY, USA, May 2014.

[116] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of information measures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 2296–2300.

[117] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[118] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[119] P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sep. 1997.

[120] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: A comprehensive study," *J. Cryptol.*, vol. 24, no. 2, pp. 269–291, 2011.

[121] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6774–6798, Oct. 2017.

[122] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1372–1376.

[123] Y. Wu and P. Yang, "Optimal entropy estimation on large alphabets via best polynomial approximation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 824–828.

[124] Y. Han, J. Jiao, and T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1367–1371.

[125] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4374–4388, Jul. 2013.

[126] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence between large-alphabet distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1118–1122.

[127] G. Valiant and P. Valiant, "Estimating the unseen: An n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, San Jose, CA, USA, Jun. 2011, pp. 685–694.

[128] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, Jun. 2016.

[129] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, Sep. 2005.

[130] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2392–2405, May 2009.

[131] X. Nguyen, M. J. Wainwright, and M. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, Nov. 2010.

[132] Y. Han, J. Jiao, and T. Weissman, "Minimax rate-optimal estimation of KL divergence between discrete distributions," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Monterey, CA, USA, Oct. 2016, pp. 256–260.

[133] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun. 2014, pp. 356–360.

[134] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.

[135] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley, 2001.

[136] D. A. Freedman, *Statistical Models: Theory and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[137] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. 21st ICML*, New York, NY, USA, 2004, pp. 29–36.

[138] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, D, J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[139] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[140] S. L. Huang and L. Zheng, "Linear information coupling problems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, USA, Jul. 2012, pp. 1029–1033.

[141] S. L. Huang and Li. Zheng, "The linear information coupling problems," 2014, *arXiv:1406.2834*. [Online]. Available: https://arxiv.org/abs/1406.2834

[142] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, Oct. 2013.

[143] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, Aug. 1969.

[144] A. Rao, A. O. Hero, III, D. J. States, and J. D. Engel, "Using directed information to build biologically relevant influence networks," *J. Bioinformat. Comput. Biol.*, vol. 6, no. 3, pp. 493–519, 2008.

[145] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal divergence estimation for finite-alphabet sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3456–3475, Aug. 2006.

[146] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weight ingmethod: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

[147] L. Zhao, Y. H. Kim, H. H. Permuter, and T. Weissman, "Universal estimation of directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 230–234.

[148] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Nov. 1990, pp. 303–305.

[149] H. Marko, "The bidirectional communication theory—A generalization of information theory," *IEEE Trans. Commun.*, vol. 21, no. 12, pp. 1335–1351, Dec. 1973.

[150] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3248–3259, Jun. 2011.

[151] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Dept. Inf. Technol. Elect. Eng., Swiss Fed. Inst. Technol., Zurich, Switzerland, 1998.

[152] Y. H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 25, no. 4, pp. 1488–1499, Apr. 2008.

[153] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 1–28, 2010.

[154] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Comput. Biol.*, vol. 2, no. 10, pp. 1301–1312, Oct. 2006.

[155] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, Jul. 2008.

[156] J. Kwon and K. M. Lee, "A unified framework for event summarization and rare event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1266–1273.

[157] I. Melnyk, A. Banerjee, B. Matthews, and N. Oza, "Semi-Markov switching vector autoregressive model-based anomaly detection in aviation systems," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 1065–1074.

[158] I. C. Paschalidis and G. Smaragdakis, "Spatio-temporal network anomaly detection by assessing deviations of empirical measures," *IEEE/ACM Trans. Netw.*, vol. 17, no. 3, pp. 685–697, Jun. 2009.

[159] L. Xiong, B. Poczos, and J. Schneider, "Group anomaly detection using flexible genre models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Granada, Spain, Dec. 2011, pp. 1071–1079.

[160] W. C. Young, J. E. Blumenstock, E. B. Fox, and T. H. McCormick, "Detecting and classifying anomalous behavior in spatiotemporal network data," in *Proc. KDD Workshop Learn. About Emergencies Social Inf. (KDD-LESI)*, New York, NY, USA, Jun. 2014, pp. 29–33.

[161] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On mining anomalous patterns in road traffic streams," in *Proc. Int. Conf. Adv. Data Mining Appl. (ADMA)*, Nanjing, China, Dec. 2011, pp. 237–251.

[162] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Bellevue, WA, USA, Nov. 2015, pp. 1–10.

[163] J. Yin, D. H. Hu, and Q. Yang, "Spatio-temporal event detection using dynamic conditional random fields," in *Proc. Int. Jont Conf. Artif. Intell. (IJCAI)*, Pasadena, CA, USA, Jul. 2009, pp. 1321–1327.

[164] A. Witayangkurn, T. Horanont, Y. Sekimoto, and R. Shibasaki, "Anomalous event detection on large-scale GPS data from mobile phones using hidden Markov model and cloud platform," in *Proc. ACM Conf. Pervasive Ubiquitous Comput. Adjunct Publication (UbiComp)*, Zurich, Switzerland, Sep. 2013, pp. 1219–1228.

[165] K. Hsiao, K. S. Xu, J. Calder, and A. O. Hero, "Multicriteria similarity-based anomaly detection using Pareto depth analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1307–1321, Jun. 2016.

[166] M. Steiger, J. Bernard, S. Mittelstädt, H. Lücke-Tieke, D. Keim, T. May, and J. Kohlhammer, "Visual analysis of time-series similarities for anomaly detection in sensor networks," *Comput. Graph. Forum*, vol. 33, no. 3, pp. 401–410, Jun. 2014.

[167] X. Li, Z. Li, J. Han, and J.-G. Lee, "Temporal outlier detection in vehicle traffic data," in *Proc. IEEE 25th Int. Conf. Data Eng.*, Apr. 2009, pp. 1319–1322.

[168] J. Xu, S. Denman, C. Fookes, and S. Sridharan, "Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach," *Expert Syst. Appl.*, vol. 54, pp. 13–28, Jul. 2016.

[169] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[170] A. Gharaibeh, M. A. Salahuddin, S. J. Hussini, A. Khreishah, I. Khalil, M. Guizani, and A. Al-Fuqaha, "Smart cities: A survey on data management, security, and enabling technologies," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2456–2501, 4th Quart., 2017.

[171] V. G. Font, C. Garrigues, and H. R. Pous, "An architecture for the analysis and detection of anomalies in smart city WSNs," in *Proc. IEEE 1st Int. Smart Cities Conf. (ISC2)*, Guadalajara, Mexico, Oct. 2015, pp. 1–6.

[172] L. Hu and Q. Ni, "IoT-driven automated object detection algorithm for urban surveillance systems in smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 747–754, May 2017.

[173] H. Zhang, Y. Zheng, and Y. Yu, "Detecting urban anomalies using multiple spatio-temporal data sources," *ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 54, Mar. 2018.

[174] Q. Wang, W. Lv, and B. Du, "Spatio-temporal anomaly detection in traffic data," in *Proc. 2nd Int. Symp. Comput. Sci. Intell. Control (ISCSIC)*, Stockholm, Sweden, Sep. 2018, pp. 1–5.

[175] X. Wu, Y. Dong, C. Huang, J. Xu, D. Wang, and N. V. Chawla, "UAPD: Predicting urban anomalies from spatial-temporal data," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, Sep. 2017, pp. 622–638.

[176] L. Hong, Y. Zheng, D. Yung, J. Shang, and L. Zou, "Detecting urban black holes based on human mobility data," in *Proc. SIGSPATIAL*, Bellevue, WA, USA, Nov. 2015.

[177] L. Jin, Z. Feng, and L. Feng, "A context-aware collaborative filtering approach for urban black holes detection," in *Proc. CIKM*, Indianapolis, IN, USA, Oct. 2016, pp. 2137–2142.

[178] V. Satuluri and S. Parthasarathy, "Scalable graph clustering using stochastic flows: Applications to community discovery," in *Proc. KDD*, 2009, pp. 737–746.

[179] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB*, vol. 2, no. 1, pp. 718–729, Aug. 2009.

[180] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 172–188, Feb. 2008.

[181] Z. Liu, J. X. Yu, Y. Ke, X. Lin, and L. Chen, "Spotting significant changing subgraphs in evolving graphs," in *Proc. ICDM*, 2008, pp. 917–922.

[182] C. Robardet, "Constraint-based pattern mining in dynamic graphs," in *Proc. ICDM*, 2009, pp. 950–955.

[183] C. S. Jensen, D. Lin, B. C. Ooi, and R. Zhang, "Effective density queries on continuously moving objects," in *Proc. ICDE*, 2006, p. 71.

[184] Z. Li, B. Ding, J. Han, and R. Kays, "Swarm: Mining relaxed temporal moving object clusters," *Proc. PVLDB*, vol. 3, nos. 1–2, pp. 723–734, 2010.

[185] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang, "On discovery of gathering patterns from trajectories," in *Proc. ICDE*, 2013.

[186] M. Mathioudakis, N. Bansal, and N. Koudas, "Identifying, attributing and describing spatial bursts," *Proc. PVLDB*, vol. 3, nos. 1–2, pp. 1091–1102, 2010.

[187] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras, "On the spatiotemporal burstiness of terms," *Proc. PVLDB*, vol. 5, no. 9, pp. 836–847, 2012.

[188] W. Song, Y. Yang, M. Fu, F. Qiu, and M. Wang, "Real-time obstacles detection and status classification for collision warning in a vehicle active safety system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 99, pp. 1–16, May 2017.

[189] H. S. Sawhney, "3D geometry from planar parallax," in *Proc. CVPR*, 1994, pp. 929–934.

[190] R. Labayrade, D. Aubert, and J. P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through 'v-disparity' representation," in *Proc. IV Symp.*, 2002, pp. 646–651.

[191] S. Kramm and A. Bensrhair, "Obstacle detection using sparse stereovision and clustering techniques," in *Proc. IV Symp.*, 2012, pp. 760–765.

[192] Z. Zhang, R. Weiss, and A. R. Hanson, "Obstacle detection based on qualitative and quantitative 3D reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 1, pp. 15–26, Jan. 1997.

[193] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: Detecting small road hazards for self-driving vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2016, pp. 1099–1106.

[194] S. K. Gehrig, F. Eberli, and T. Meyer, "A real-time low-power stereo vision engine using semi-global matching," in *Proc. 7th Int. Conf. Comput. Vis. Syst.*, vol. 5815, 2009, pp. 134–143.

[195] H. Badino and T. Kanade, "A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion," in *Proc. MVA*, 2011, pp. 185–189.

[196] A. Petrovskaya and S. Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Auton. Robots*, vol. 26, nos. 2–3, pp. 123–139, Apr. 2009.

[197] Y. Ye, L. Fu, and B. Li, "Object detection and tracking using multi-layer laser for autonomous urban driving," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 259–264.

[198] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "RefineNet: Refining object detectors for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 1, no. 4, pp. 358–368, Dec. 2016.

[199] A. Unterholzner and H. Wuensche, "Selective attention for detection and tracking of road-networks in autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Gold Coast, QLD, Australia, Jun. 2013, pp. 277–284.

[200] A. J. Davison and D. W. Murray, "Simultaneous localization and map-building using active vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 865–880, Jul. 2002.

[201] C. Premachandra, H. Waruna, H. Premachandra, and C. D. Parape, "Image based automatic road surface crack detection for achieving smooth driving on deformed roads," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Manchester, U.K., Oct. 2013, pp. 4018–4023.

[202] H. Oliveira and P. L. Correia, "Automatic crack detection on road imagery using anisotropic diffusion and regional linkage," in *Proc. 18th Eur. Signal Process. Conf.*, 2010, pp. 274–278.

[203] M. A. Sotelo, F. J. Rodriguez, and L. Magdalena, "VIRTUOUS: Vision-based road transportation for unmanned operation on urban-like scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 2, pp. 69–83, Jun. 2004.

[204] M. A. Sotelo, F. J. Rodriguez, L. Magdalena, L. M. Bergasa, and L. Boquete, "A color vision-based lane tracking system for autonomous driving on unmarked roads," *Auto. Robots*, vol. 16, no. 1, pp. 95–116, Jan. 2004.

[205] C. Fernández, R. Izquierdo, D. F. Llorca, and M. A. Sotelo, "Road curb and lanes detection for autonomous driving on urban scenarios," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 1964–1969.

[206] D. Campo, M. Baydoun, P. Marin, D. Martin, L. Marcenaro, A. de la Escalera, and C. Regazzoni, "Learning probabilistic awareness models for detecting abnormalities in vehicle motions," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[207] G. Wang, J. Guo, Y. Chen, Y. Li, and Q. Xu, "A PSO and BFO-based learning strategy applied to faster R-CNN for object detection in autonomous driving," *IEEE Access*, vol. 7, pp. 18840–18859, Feb. 2019.

[208] X. Z. Chen, K. Kundu, Y. Zhu, S. Fidle, R. Urtasun, and H. Ma, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, May 2018.

[209] T. Yu, X. Wang, and A. Shami, "Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2207–2216, Dec. 2017.

[210] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.

[211] J. A. Stankovic, "Research directions for the Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 3–9, Feb. 2014.

[212] Y. Liu and Q. Wu, "A lightweight anomaly mining algorithm in the Internet of Things," in *Proc. 5th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Jun. 2014, pp. 1142–1145.

[213] V. A. Desnitsky, I. V. Kotenko, and S. B. Nogin, "Detection of anomalies in data for monitoring of security components in the Internet of Things," in *Proc. 18th Int. Conf. Soft Comput. Meas. (SCM)*, St. Petersburg, Russia, May 2015, pp. 189–192.

[214] D. Stiawan, M. Y. Idris, R. F. Malik, S. Nurmaini, and R. Budiarto, "Anomaly detection and monitoring in Internet of Things communication," in *Proc. 8th Int. Conf. Inf. Technol. Elect. Eng. (ICITEE)*, Yogyakarta, Indonesia, Oct. 2016, pp. 1–4.

[215] J. Takahashi, D. Shioiri, Y. Shida, Y. Kobana, R. Suzuki, Y. Kobayashi, N. Isoyama, G. Lopez, and Y. Tobe, "Clustering for road damage locations obtained by smartphone accelerometers," in *Proc. 2nd Int. Conf. IoT Urban Space (UrbIoT)*, Tokyo, Japan, May 2016, pp. 89–91.

[216] J. Borges, T. Riedel, and M. Beigl, "Urban anomaly detection: A use-case for participatory infra-structure monitoring," in *Proc. 2nd Int. Conf. IoT Urban Space (UrbIoT)*, Tokyo, Japan, May 2016, pp. 36–38.

[217] M. Shahriar and M. Rahman, "Urban sensing and smart home energy optimisations: A machine learning approach," in *Proc. Int. Workshop Internet Things Towards Appl. (IoT-App)*, Seoul, South Korea, 2015, pp. 19–22.

[218] N. Rakesh, "Performance analysis of anomaly detection of different IoT datasets using cloud micro services," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Coimbatore, India, Aug. 2016, pp. 1–5.

[219] S. Brady, A. Hava, P. Perry, J. Murphy, D. Magoni, and A. Dominguez, "Towards an emulated IoT test environment for anomaly detection using NEMU," in *Proc. Global Internet Things Summit (GIoTS)*, Geneva, Switzerland, Jun. 2017, pp. 1–6.

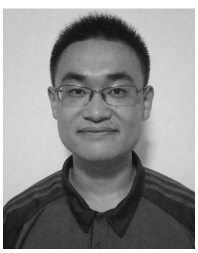
[220] I. Nevat, D. M. Divakaran, S. G. Nagarajan, P. Zhang, L. Su, L. L. Ko, and V. L. L. Thing, "Anomaly detection and attribution in networks with temporally correlated traffic," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 131–144, Feb. 2018.

[221] N. Moustafa, K. R. Choo, I. Radwan, and S. Camtepe, "Outlier Dirichlet mixture mechanism: Adversarial statistical learning for anomaly detection in the fog," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 1975–1987, Aug. 2019.

[222] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in Internet of Things," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4672–4683, Jun. 2019.

RUI SHE received the B.S. degree in communication engineering from Jilin University, Changchun, China, in 2015. He is currently pursuing the Ph.D. degree in electronic engineering with Tsinghua University, Beijing, China. From December 2018 to June 2019, he was a Visiting Scholar with Columbia University, New York, NY, USA. His research interests include data mining, machine learning, and big data analysis with information theory.

SHANYUN LIU received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include wireless communications in big data analysis, high speed railway, and information theory.



SHUO WAN received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include information theory and multi-robot systems.

KE XIONG (M'14) received the B.S. and Ph.D. degrees from Beijing Jiaotong University (BJTU), Beijing, China, in 2004 and 2010, respectively.

From April 2010 to February 2013, he was a Postdoctoral Research Fellow with the Department of Electrical Engineering, Tsinghua University, Beijing. Since March 2013, he has been a Lecturer and an Associate Professor with BJTU. From September 2015 to September 2016, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. He is currently a Full Professor with the School of Computer and Information Technology, BJTU. He has published more than 100 academic papers in referred journals and conferences. His current research interests include wireless cooperative networks, wireless powered networks, and network information theory. He is a member of the China Computer Federation (CCF) and a Senior Member of the Chinese Institute of Electronics (CIE). He is also as an Associate Editor-in-Chief of the Chinese journal New Industrialization Strategy and an Editor of Computer Engineering and Software. In 2017, he has served as a leading editor of the Special issue Recent Advances in Wireless Powered Communication Networks for the *EURASIP Journal on Wireless Communications and Networking* and a Guest Editor of the Special issue Recent Advances in Cloud-Aware Mobile Fog Computing for *Wireless Communications and Mobile Computing*. He also serves as a Reviewer for more than 15 international journals, including the IEEE Transactions on Signal Processing, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Communications, the IEEE Transactions on Vehicular Technology, the IEEE Communication Letters, the IEEE Signal Processing Letters, and the IEEE Wireless Communication Letters. He has also served as a Session Chair for the IEEE GLOBECOM2012, IET ICWMMN2013, the IEEE ICC2013, ACM MOMM2014, and the Publicity and Publication Chair for the IEEE HMWC2014, as well as the TPC Co-Chair of IET ICWMMN2017.

PINGYI FAN (M'03–SM'09) received the B.S. degree from the Department of Mathematics, Hebei University, in 1985, the M.S. degree from Nankai University, in 1990, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1994.

From August 1997 to March 1998, he visited the Hong Kong University of Science and Technology as a Research Associate. From May 1998 to October 1999, he visited the University of Delaware, Newark, DE, USA, as a Research Fellow. In March 2005, he visited NICT, Japan, as a Visiting Professor. From June 2005 to May 2017, he visited the Hong Kong University of Science and Technology for many times. From July 2011 to September 2011, he was a Visiting Professor of the Institute of Network Coding, Chinese University of Hong Kong. He is currently a Professor with the Department of EE, Tsinghua University. His main research interests include big data analytics, machine learning, 5G technology in wireless communications, such as massive MIMO, OFDMA, network coding, and network information theory. He is an Oversea Member of IEICE. He has attended to organize many international conferences, including as the General Co-Chair of the IEEE VTS HMWC2014, the TPC Co-Chair of the IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS 2010), and a TPC Member of the IEEE ICC, Globecom, WCNC, VTC, and Infocom. He has received some academic awards, including the IEEE Globecom 2014 Best Paper Award, the IEEE WCNC'08 Best Paper Award, the ACM IWCMC'10 Best Paper Award, and the IEEE ComSoc Excellent Editor Award for the IEEE Transactions on Wireless Communications, in 2009. He has served as an Editor of the IEEE Transactions on Wireless Communications, the *Inderscience International Journal of Ad Hoc and Ubiquitous Computing*, and the *Wiley Journal of Wireless Communications and Mobile Computing*. He is also a Reviewer of more than 32 international journals, including 20 IEEE journals.