International Journal of Data Warehousing and Mining
Volume 19 • Issue 1

A Survey of Collective Anomaly Detection on Sequence Dataset

Xiaodi Huang, Hefei University, China*
https://orcid.org/0000-0001-6530-2462

Po Yun, Hefei University, China

Zhongfeng Hu, Hefei University, China

ABSTRACT

Anomaly detection on sequence dataset typically focuses on the detection of collective anomalies, aiming to find anomalous patterns consisting of sequences of data with specific relationships rather than individual observations. In this survey, existing studies are summarized to align with temporal sequence dataset and spatial sequence dataset. For the first category, the detection can be subdivided into symbolic dataset based and time series dataset based, which include similarity, probabilistic, and trend approaches. For the second category, it can be subdivided into homogeneous datasets based heterogeneous datasets based, which include multi-dataset fusion and joint approaches. Compared to the state-of-the-art survey papers, the contribution of this paper lies in providing a deep analysis of various representations of collective anomaly in different application field and their corresponding detection methods, representative techniques. As a result, practitioners can receive some guidance for selecting the most suitable methods for their particular case.

KEYWORDS

Collective Anomaly, Data Relation, Sequence Dataset, Spatial Sequence Dataset, Temporal Sequence Dataset

1. INTRODUCTION

With the continuous development of the IoT, interconnected sensing devices can collect and transfer a large amount of data from various application fields among themselves. These data vary greatly in structure and correlation, but most of them are generated in the form of sequence (Chowdhury, 2019). The sequence dataset is an ordered list made up of sequential items, such as events or numbers, on which there is a strong correlation among them. How to effectively mine such datasets is a hot topic within various disciplines, areas, and applications (Truong, Hai, Le, Fournier-Viger, & Fujita, 2021). In addition to common research based on frequent pattern mining, anomaly detection research that focuses on rare pattern mining is also becoming increasingly popular.

DOI: 10.4018/IJDWM.327363

*Corresponding Author

1.1 Research Background

Anomaly detection, as an analytical approach to data structure, aims to exhibit implicit knowledge by mining rare patterns that do not satisfy the overall expectations. It is widely used in many application fields, such as fraud detection (Trivedi, Monik, & Mridushi, 2016), intrusion prevention (R. Zhang, Xia, Shao, Ren, & Cheng, 2020), image identification (K. Zhang, Wang, & Kuo, 2022), etc. Currently, the meaning of an anomaly is commonly referred to by the following definition, which was originally proposed by Hawkins (Auth, 1980): "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.' Anomaly and outlier are interchangeable terms commonly used in this field.

As a data pattern, the anomaly is similar to novelty but distinct from noisy data (Cappozzo, Greselin, & Murphy, 2020). An anomaly is not necessarily incorrect by itself but contains implicit information, which is always more valuable for analysis than normal forms. Noisy data are usually generated by observation errors such as missing data or random variance, such as acquisition errors, hardware fault, which are not produced by any data generation mechanism. Not only are they worthless, but they also act as irritating hindrances to analysis. Novelty data are mainly associated with evolving datasets, for example, social evolution and natural sciences, which represent an unknown data pattern that may reflect new themes (Shah, Azam, Ali, Khan, & Yao, 2021). In the initial stages, novelty can be treated as an anomaly, and the discovery approach is similar to that of anomaly detection. The main difference is that each novelty will be incorporated into the normal pattern after verification, and the following data that belong to it will no longer be treated as novelty. Whereas, no matter how many times the anomaly is recorded, if the conditions remain unchanged, it will still be judged as an anomaly in the next occurrence.

The desired trait of the anomaly is the key point to selecting an appropriate anomaly detection approach (Al-Ghaili, Ibrahim, Hairi, Rahim, & Kasim, 2021). Based on it, the anomaly can be divided into point anomaly, contextual anomaly and collective anomaly. Detection approaches based on contextual and point anomaly mainly focus on analyzing whether a single data presents abnormal performance. If the contextual attributes are empty, point anomaly is a special type of contextual anomaly (Zamini & Hasheminejad, 2019). For sequence datasets, most abnormal behaviors are characterized by complexity and interactivity, which are difficult to detect by a single data observation itself or a single interval, but can be identified as collective anomaly(Huiling Qin, Xianyuan Zhan, & Yu Zheng, 2021).Thus, as a special category, collective anomaly is used to reflect abnormal patterns which are presented by multiple segments or datasets based on association relations. Therefore, the detection methods and techniques should be flexible for different applications.

1.2 Related Work

With the gradual deepening of anomaly detection research, aiming to exhibit a clearer perspective to interested readers, researchers have summarized existing achievements from the perspective of anomaly notion, detection method, and application field.

In the first survey of anomaly detection (Chandola, Banerjee, & Kumar, 2009), based on the data characteristic of various anomalies, the detection method including mechanism and technique are classified and introduced in detail corresponding to each application fields, such as intrusion, fraud, image processing, and text analysis. Although the survey first proposed the concept of collective anomaly based on the nature of the data, the classification still considers the anomaly as a whole rather than providing a targeted analysis of the collective anomaly. This survey lays the foundation for future more targeted investigations and demonstrates the urgent need for intensified research in this field. Prasanta, Bhattacharyya, Borah, and Kalita (2011) classified existing detection methods according to metric data, evolving data, and multi-structured data. This research clearly explains the relationships between different detection methods and data structures, but the detection performance of different types of anomalies varies in similar data structures. Therefore, to improve the detection performance, Chouhan, Richhariya, Chouhan, and Richhariya (2015) considered that the existing classical detection methods should be adjusted due to the characteristics of the data presented by point anomaly, contextual anomaly, and collective anomaly.

Although these three types of anomalies have their own unique features and areas, there are many overlaps among them. If the collective attribute is regarded as an empty set, the point anomaly is a special kind of collective anomaly. If the behavior attribute and the contextual attribute are regarded as two kinds of collective attribute, the contextual anomaly can also be regarded as a special collective anomaly. Therefore, collective anomaly detection is at a higher level in the study of anomaly detection, and a comprehensive study of its existing detection methods, detection techniques, and application fields can further improve the precision of research in the field of anomaly detection.

For collective anomaly, the generation mechanism determines that it can only occur in the dataset where the data are related, hence, the effectiveness of detection method is greatly affected by the data structure. According to the structural feature of graph dataset, Feroze, Daud, Amjad, and Hayat (2021) classified the graphical detection methods of collective anomaly as static versus dynamic and attributed versus plain graph methods, which significantly improves the detection effectiveness in specific application fields like deep learning and machine learning. However, the detection methods of collective anomaly are significant differences in various data structures, even the methods belong to the same type also should be adjusted according to the structural feature. Therefore, more discriminant approaches need to be developed with more discriminatory anomaly detection ability.

At present, increasingly studies have shown that a considerable portion of anomaly detection on sequence datasets are occupied by collective anomaly, and this proportion is much larger than that appearing in other forms of datasets. Therefore, we also take the sequence dataset as background, aiming to provide a comprehensive overview and exploration directions for readers interested in this research area.

This survey aims to make such a comprehensive analysis of collective anomaly detection based on different data relations that exist in different types of sequence datasets. Existing studies mostly focused on the general concept of collective anomaly detection, but few attempts were made to analyze the multimodal data relations of anomaly in different types of sequence datasets. The different relations will lead to different anomalous patterns, which may require specific detection approaches. For various applications, detection approaches of collective anomaly may vary considerably in methods and techniques. Even if a detection approach is evaluated in the targeted application, it can hardly guarantee its effect in other fields. According to the various data relations, their sequential traits must be considered for both the construction of the definition of anomaly and the selection of the detection approach. For different applications, how to identify and model these data relations is essential for conducting an effective detection.

The specific contributions of this survey are as following:

1. Based on existing research on collective anomaly, we extend the meaning of ‘collective’ to represent more than just a collection of data. For a temporal sequence dataset, the detection is conducted in a single dataset, and collective anomaly represents a collection during a few consecutive time intervals, while a single subsequence in the collection may not be anomalous at a single time interval if being checked individually. For spatial sequence dataset, collective anomaly might not be that anomalous in terms of a single dataset but considered an anomaly when checking multiple datasets simultaneously.
2. Existing detection approaches are summarized to align with the temporal sequence dataset and the spatial sequence dataset, and a more detailed classification is divided according to their marked different data structure. For different types of sequence datasets, due to the relations among the data, collective anomaly that appears in them may present as an entire sequence, a dependent subsequence, or a set of various subsequences. For example, given a database of scientific experiment data, one might be interested in sequences that are anomalous. On the other hand, given a long sequence of mechanical data from the gearbox, one might be interested in

detecting subsequences when the malfunction appears. Even though their corresponding detection approaches are similar in principle, it is still different to achieve interoperability. A wide range of scattered application research will confuse the readers, and a systematic survey is urgently needed to sort these diverse studies.

3. A clear classification for existing research of collective anomaly models based on different types of sequence datasets is essential to widespread application. The complexity of computation will be a fatal limitation when dealing with high dimensional datasets. The characteristics of sequential data can be very complex, which may be reflected in the size of data string, the number of data attributes, and the structure of data relations. To achieve the goal of dimension reduction, in addition to optimizing the detection technique, it is also necessary to construct a suitable model to identify anomalous patterns that constitute collective anomalies.

1.3 Paper Organization

The rest of this survey are organized as follows. In Section 2, the authors compared three anomaly patterns to highlight the particularity of collective anomaly. Moreover, on the different relations among datasets, sequence datasets are classified as temporal and spatial sequence datasets. Additionally, according to their various data structures, each of them has been subdivided into two categories. In Section 3, focusing on the temporal sequence dataset, we analyze the two most common data structures on such datasets in detail: the symbol dataset and time series dataset. Based on their distinct differences, we have proposed detection methods and techniques for each of them separately. In Section 4, focusing on the spatial sequence dataset, due to the different interaction patterns among multi-datasets, the spatial datasets can be classified as homogeneous dataset and heterogeneous dataset. For each of them, we propose the detection approaches separately. In the last section, we summarize the contributions of this paper and proposed several important ideas for future research directions.

2. RESEARCH FOUNDATION

2.1 Detection Framework

The collective anomaly detection process can be divided into a three-step framework as shown in Figure 1.

For each input sequence dataset, whether they belong to temporal sequence datasets or spatial sequence datasets should first be determined based on the data attributes. Then, based on their different data relations, the temporal datasets can be subdivided as symbolic and time series datasets, while spatial datasets can be subdivided as homogeneous and heterogeneous datasets. Finally, based on the classification in the previous steps, the detection approach, including method, model and technique, should be chosen according to the anomaly patterns of each sequence dataset.

2.2 The Taxonomy of Anomaly

The trait of expected anomaly is an important factor while considering the detection approach. As mentioned above, it can be classified as point anomaly, contextual anomaly, and collective anomaly.

2.2.1 Point Anomaly

In a given dataset, observations will be termed a point anomaly if they deviate significantly from the others on the metric of target attributes (Q. Liu et al., 2017). This is the simplest type of anomaly, and most of the current research aims to find such an anomaly. In practical application, the key to detecting it is to find the appropriate attributes for the deviation metric.

Figure 2 shows a dataset that collects the response of active volcano movement to seasonal ocean currents through specific heat capacity (C, J/kg●°C). The values at o_1 and o_2 are typical point anomalies, while their values are abnormal from the historical average.

Figure 1. The framework of collective anomaly detection

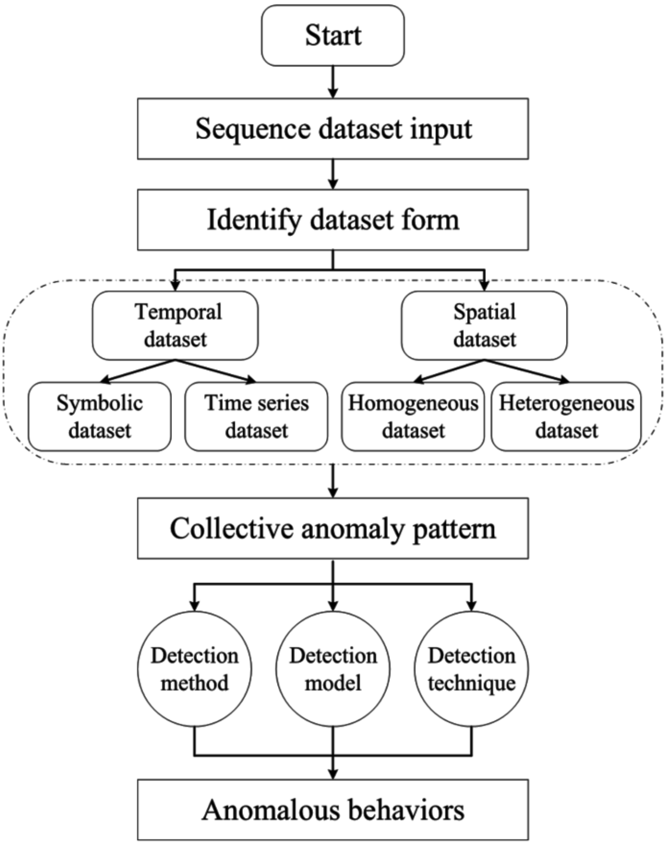
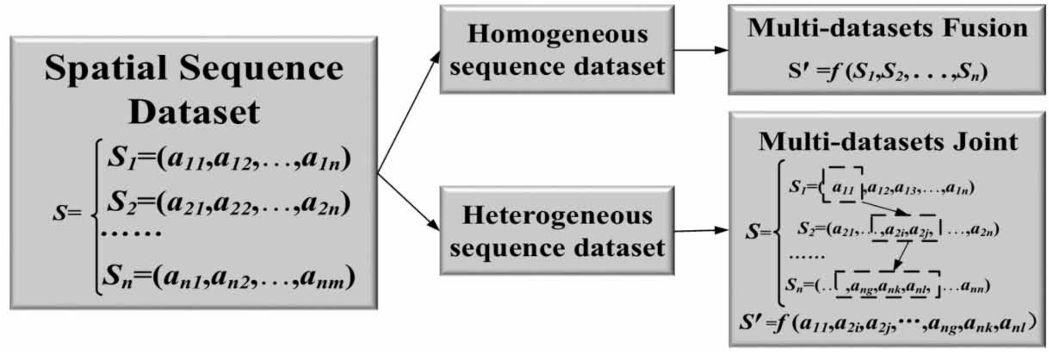


Figure 2. Point anomaly o_1 and o_2



2.2.2 Contextual Anomaly

In a given dataset, observation will be identified as contextual anomaly, if it deviates significantly from a normal pattern under a particular context but not otherwise (Golmohammadi & Zaiane, 2017).

Figure 3 shows a monitoring dataset of submarine active volcano movement, the specific heat capacity is a behavioral attribute used to reflect the stability of volcanic by monitoring the state of seawater. Both point o_1 and subsequence O_1 are contextual anomalies, although the values are within the reasonable ranges, their behavior patterns are significantly deviating from normal expectations during their collected timestamp. The context must be taken into account for the detection.

In such scenarios, each data can be divided into behavioral and contextual attributes. The behavioral attributes of the collected observations are usually composed of a series of objective measurements that are used to describe the physical characteristics of the target problem.

Contextual attributes are used to reflect the background information of the dataset. It cannot describe the physical properties of the problem independently but must be analyzed in combination with the behavioral attributes. For an observation, even if it is a contextual anomaly in the given context, the identical observation (in terms of behavioral attributes) may be considered normal in another context. In Figure 3, time is the contextual attribute that reflects the month of monitoring data acquisition. At time point and time interval [t_i: t_j]. Without the contextual attribute, these observations are within a reasonable range, but they will be identified as contextual anomalies in terms of their collected timestamps. In practice, for such anomalies, how to choose a meaningful contextual attribute is the main factor when considering the selection of the detection approach.

2.2.3 Collective Anomaly

The collective anomaly is a set of related data. When they appear together in a certain pattern, their overall behavior attribute will deviate significantly from the entire dataset, but the individual data may not be an anomaly (Chandola et al., 2009).

In Figure 4, the subsequence O_1 denotes a collective anomaly, since the variation trend of the subsequence is distinct different from the historical curve during the time interval, even if each observation fluctuates within the normal range by itself.

A single data observation could be generated by arbitrary mechanisms, it is not enough to judge whether a collective anomaly occurs by examining a single data observation. For example, the same stock to be heavily traded between two share-holders may be considered normal, but if a large number of the same stock are repeatedly traded among a small group of shareholders, it will be treated as a collective anomaly. This is probably illegal manipulation of the stock market, such as

Figure 3. Contextual anomaly o_1 and O_1

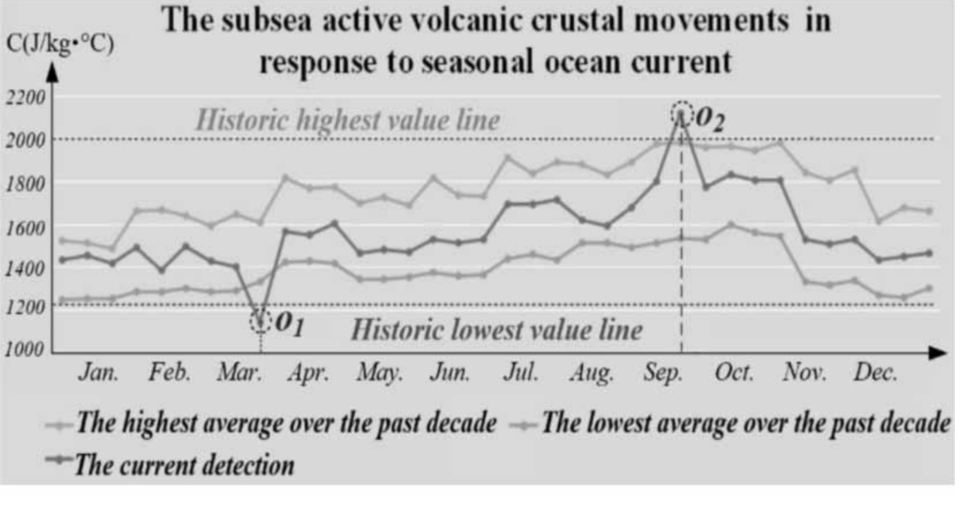
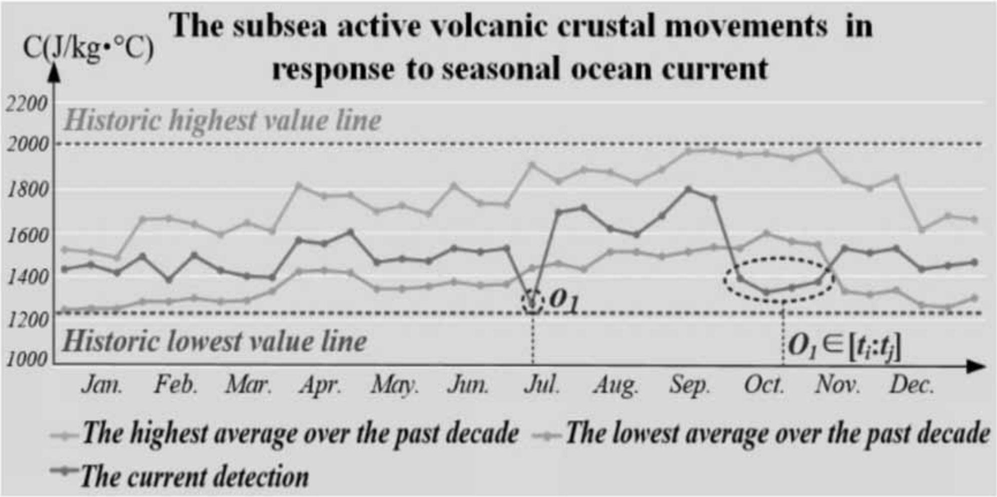


Figure 4. Collective anomaly O_1



money laundering. Therefore, to detect such an anomaly, a relational model based on data attributes must be built in the neighborhood of the data. It intuitively means that collective anomaly can only occur in the dataset where the data observations are related.

2.3 Sequence Dataset

Refer to its generating mechanism, the sequence dataset is always collected as an ordered list of data observations or event sets. Currently, sequential data is increasingly common in a wide variety of application fields, e.g., network traffic (Yuan, Xiong, & Wang, 2019), gene research (Matsushita, Egami, Sawada, Saito, & Yoshida, 2019) and fraud detection (Jurgovsky et al., 2018), etc.

As mentioned above, for collective anomaly, data relations represented by data attributes must be taken into account when building the detection model. Therefore, in the face of a wide variety of specific problems, it is almost always impossible to apply a general model. Even if the detection model is effective in an application scenario, it may be difficult to adapt directly to others. Detection models for collective anomaly need to be redesigned for specific problems. This is mainly because both the data attributes and the definitions of collective anomaly are distinct in different types of sequence dataset. To sort out existing papers and form a coherent framework, the authors divide sequence datasets into temporal and spatial datasets due to their relations.

2.3.1 Temporal Sequence Dataset

The temporal sequence dataset is a set of data observations that must be analyzed with their collected timestamps (Dasgupta, Yoshizumi, & Osogami, 2016). Each data observation can be abstracted as a binary group: <*b*, *t*>, *b* is the behavioral attribute, *t* is the temporal attribute. For a temporal sequence dataset, the behavioral attribute of each observation is described by a finite number of features, according to different detection targets, the appropriate combination of features will be selected. The analysis of such application is always conducted in a single dataset, there is no certain relation between different datasets.

Depending on the degree of temporal attributes participate in the analysis, it can be subdivided into two categories: symbolic sequence and time series dataset.

Symbolic sequence dataset consists of an ordered set of events which are represented by a limited size of letters with no sequential meaning or other types of nominal data (Xin & Xu, 2011).

Its timestamp is just a statement to indicate the order in which the data observation is collected. The value of each timestamp does not participate in the analysis. Applications such as project management (Deng, Wang, Zhao, Zou, & Chen, 2021), customer shopping (Y. Liu, Guo, Li, Zhang, & Yao, 2019), web click flow (Yin, Zhang, & Kwang-Jun, 2017) involve such datasets.

Time series dataset commonly consist of long or even infinite sequences, in which data observations are always collected at equal or continuous time intervals (Valencia, Astray, Fernández-González, Aira, & Rodríguez-Rajo, 2019). The behavioral attributes of each observation correspond to their temporal attributes, the value of each timestamp must be recorded and taken into account in the analysis. Such sequences can be easily found in applications where activities have been supervised during a long period of time, e.g., statistics, monitoring, and so on.

2.3.2 Spatial Sequence Dataset

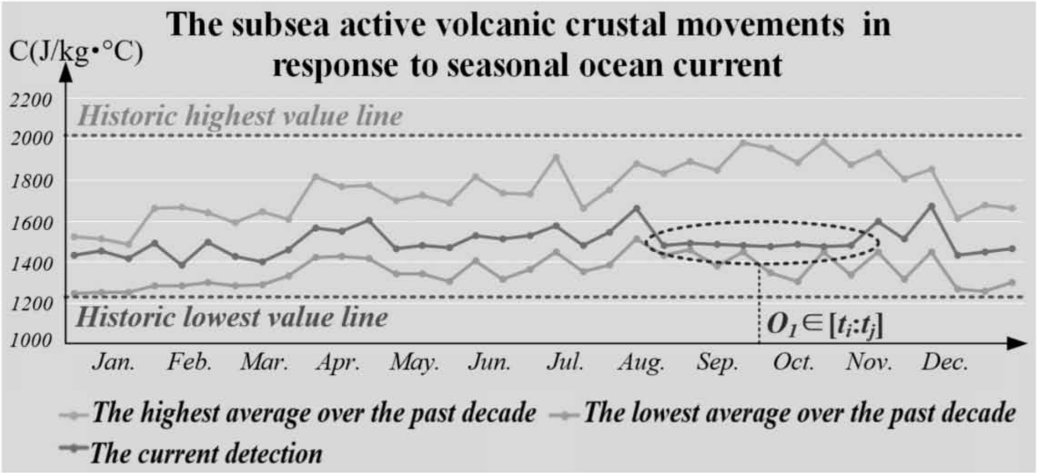
As an extension of the temporal sequence dataset, each data observation that consists of the spatial sequence dataset can be abstracted into a triple: <*b*, *t*, *p*>, where *p* is the position tag used to reflect the spatial attribute. The spatial attribute of a dataset is usually determined by its generation mechanism, such as physical location and logical relations (Liao, Hou, & Jiang, 2019). Unlike the temporal dataset, the detection target related to spatial sequence datasets cannot be solved by mining a single dataset mining. The spatial attribute should be used as the medium to analyze the association between multiple datasets, collective anomaly detection based on spatial sequence datasets is usually based on joint mining of datasets. Spatial sequence datasets can often be found in complex decision making areas, such as combat planning and strategy formulation in military fields (Wen, Sun, Zeng, Zhang, & Yuan, 2018), road network planning and optimization in traffic fields (Zhifeng et al., 2017), species and pollutant tracking in ecological fields (Ghosh, Li, Ca O, & Ramamohanarao, 2017).

Due to the different generation mechanisms and relations among multiple spatial datasets, they can be divided into heterogeneous and homogeneous datasets.

If datasets generated at different positions are used to describe similar behavioral attributes that aim to represent the same target problem from different perspectives, they will be treated as homogeneous datasets. We assume that each of them enjoys similar data attributes and structures.

As shown in Figure 5, the urban traffic dataset is composed of private car traffic, bus traffic, and rental bike traffic in six regions from r_1 to r_6, which can be regarded as homogeneous datasets. According to the traffic information of different vehicles, each dataset reflects the traffic condition

Figure 5. Homogeneous datasets reflected by urban traffic datasets



from different angles. Thus, it can be assumed that these datasets have similar data structures and attributes.

If datasets are used to describe different behavioral attributes of various problems, while these problems are interrelated, they can be treated as heterogeneous datasets. The data that make up these datasets are not identical in categories and distributions.

For example, in large manufacturing industries, the set of equipment is processed and assembled through different links. These datasets collected at various links are used to reflect the quality information of each component. There is significant distinguishing on the quality control indicators and scope of qualification, and observations that constitute these datasets vary in collection interval and categories. For heterogeneous dataset that formed by these data observations, it is not enough to judge whether the whole equipment is qualified even if a single part meets the standard, the dataset of each link must be analyzed jointly.

Different types of sequence datasets will result in different data structures and attributes, which will have an important impact on the definition of collective anomaly. Therefore, the key to detecting collective anomaly in a sequence dataset is to build an appropriate model that adequately reflects the different relations. Different collective anomaly definitions and their detection mechanisms according to various types of sequence datasets will be discussed in Sections 3 and 4. This paper was not only to determine which problems could be solved by a certain technique, but also to build several general detection models to solve these different but similar problems.

3. DETECTION ON TEMPORAL SEQUENCE DATASET

As mentioned above, the temporal sequence dataset can be subdivided into a symbolic dataset and a time series dataset. In this section, we first analyze collective anomaly patterns that might emerge in each of them. Then, detection models are built according to each pattern, and the typical techniques and their variations for each model are given.

3.1 Symbolic Sequence Dataset

The symbol sequence dataset is usually made up of nominal data observations that are used to describe the attributes of the category without giving actual values (Marquez-Grajales, Acosta-Mesa, & Mezura-Montes, 2017). Each data represents a certain state, category, or encoding, for which they have little mathematical meaning. Even if a numerical value is given, it is used as a code to represent a certain class, but not for numeric calculation.

In the field of computer security, the subsequence highlighted in Figure 6 reflects the computer intrusion based on web attack, and the subsequence highlighted in Figure 7 is based on host intrusion. In these sequences, data observations that make up them seem normal in themselves, but when they occur as a set or in a particular order, the sequence formed from them will be treated as a collective anomaly.

For such sequence datasets, collective anomaly patterns are known or predictable. The detection approach can be divided into two subtasks: mining patterns, such as combination (shown in Figure 6) or co-occurrence (shown in Figure 7), and generating the data relations. As shown in Figure 8, the detection method aims to identify the abnormal subsequence which is different from the expectation, even if each data that constitute the subsequence are normal by itself.

3.1.1 Clustering Based Approach

Symbolic sequence datasets usually have high dimensionality, complexity, and noise (Verenich, Dumas, Rosa, Maggi, & Francescomarino, 2015). In many fields, obtaining sufficient available labeled training data is costly, thus, the unsupervised based collective anomaly detection approach will be more suitable.

Figure 6. The sequence dataset of computer network acceptance

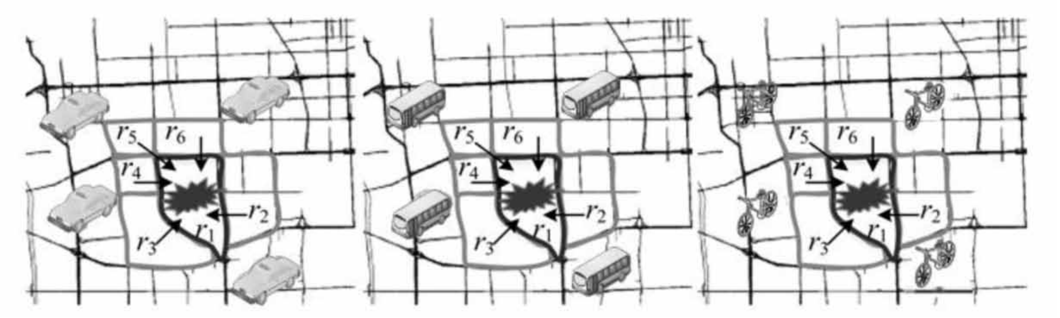


Figure 7. The sequence dataset of computer system calls

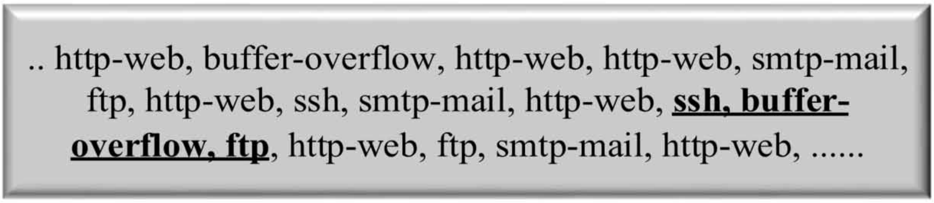
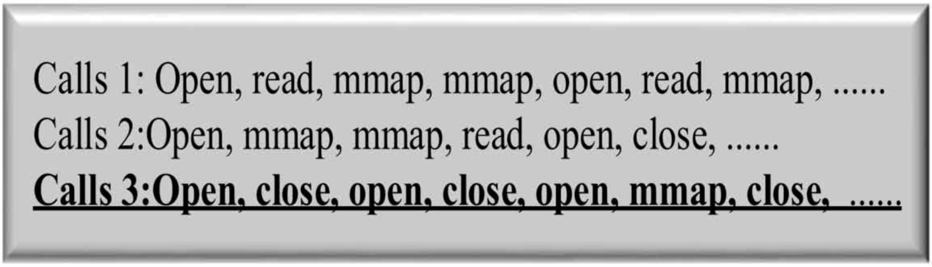

.. http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, **ssh, buffer-overflow, ftp**, http-web, ftp, smtp-mail, http-web, ...

Figure 8. The detection model of symbolic sequence dataset


Calls 1: Open, read, mmap, mmap, open, read, mmap, ...
Calls 2:Open, mmap, mmap, read, open, close, ...
Calls 3:Open, close, open, close, open, mmap, close, ...

Clustering based approach, as one of the most famous unsupervised techniques, has been widely used in collective anomaly detection of symbolic datasets. The abnormal pattern will be treated as a small cluster on the assumption that the majority of a given dataset is normal, but only a small percentage is abnormal. Depending on the target task, the clustering rules and features that can accurately identify abnormal patterns are selected. On this basis, the sequences belonging to these clusters will be treated as collective anomalies. The existing clustering techniques can be divided into two types as follows: partitioned clustering approach and hierarchical clustering approach. The efficiency of such detection is highly correlated with the validity of the clustering criteria.

In this part, the detection process of Denial of service (DoS) attack is chosen as an example to demonstrate several clustering approaches that can be applied to such situations. DoS attack is a common mode of computer intrusion based on web attack, which causes the collapse of computer systems by sending a mass of normal but meaningless calls to the target computer (Yin, Sun, Jin, & Kim, 2015). Although the DoS attack is illegal, each instruction that makes up it is legal. Considering such characteristics, the DoS attack cannot be treated as a point anomaly and in this regard, treating it as a collective anomaly will be a better idea for more accurate detection.

DoS attack detection is a kind of partitioned clustering approach, *k-means* algorithm based on distance measurement (Gomathi & Umagandhi, 2016), DBSCAN algorithm (Gokcesu & Kozat, 2017) based on density are the two most typical algorithms. Ahmed *et al*. chose the 'payload length' as the cluster feature, and the x-means clustering algorithm (Ahmed & Mahmood, 2014), which is a variant of the *k-means* algorithm, was adopted to build the clustering rules. When a group of similar data behaves anomalously with respect to the entire dataset, they will be treated as collective anomalies. In order to create a more fine grained representation of the data, Mahmood *et al.* (Ahmed & Mahmood, 2015) extended co-clustering algorithm by integrating information theory, important traffic attributes are selected as clustering rules based on mathematical logic criteria. If the volume of a set of normal computer calls is significantly different from the expectation, they will be treated as collective anomalies. Most of the current research in this area is based on sliding window sampling and optimized clustering (Lin & Su, 2019).

For DoS attack detection based on the hierarchical clustering approach, the main idea is to reflect more details of the dataset through multilevel clustering (Mohiuddin & Ahmed, 2017) (the exact number of clustering layers are set due to accuracy requirements). The abnormal pattern is diagnosed by comparing interactive information between the upper and lower levels of the clusters, for example, assuming that the upper level cluster C is $\{i, i, k, i, i, j, i, l, i\}$, the two lower levels of the clusters which can be obtained by hierarchical clustering are $M_1 = \{i, i, i, i, i\}$ and $M_2 = \{j, k, l\}$. When comparing the information within the clusters of each level, it can be found that the interaction information between the M_1 cluster and the C cluster is much more similar than that between the M_2 cluster and the C cluster. Such a large number of similar normal instructions are highly clustered, which will increase the probability of being recognized as a collective anomaly. In such studies, the focus of current research is on how to incorporate the domain knowledge of experts to handle the diversity of collective anomaly (Ding, Wang, Ge, & Li, 2018).

3.1.2 Classification-Based Approach

For the problem with low dimensionality that may have sufficient training data, the collective anomaly detection can be considered as a classification problem where it builds models of normal behavior, of which it uses to detect abnormal patterns that significantly deviate from the model (Bigdeli, Raahemi, Mohammadi, & Matwin, 2015).

For example, in the operation of an urban power grid, regardless of some faults that are unavoidable (such as natural calamities), many other types of fault usually have perceptible symptoms before complete breakdown occurs (Langarica, Ruffelmacher, & Nunez, 2019). If these faults can be detected while still under their incipient failure conditions, the loss will be minimized. The regional fault of the urban power grid is just one of such type, there is always a period between the initial occurrence of the fault and the formation of the final fault phenomenon (Liang, Ali, & Zhang, 2020).

Any single fluctuating signal in an urban power grid is regarded as a normal data instance, if a large number of fluctuating signals appear in the coverage of a base station at the same time, they will be identified as collective anomaly. Huang *et al*. (Xiaodi, Minglun, & Zhongfeng, 2020) proposed a novel detection approach based on fixed point iteration algorithm (Huang, Ren, & Hu, 2020), which transforms the diagnosis of regional fault into the detection of collective anomaly from the data of current fluctuation signal. Firstly, multi-layered classification is carried out by taking the information of upstream base stations with different energy levels corresponding to the abnormal current fluctuation signals as metrics to extract the profiling information of abnormal signals in the power grid and store them in the form of structure tree. Then, for these tree structure clusters, comparing the cluster structure information of the same layer with that of the upper and lower layer according to different judgment rules to identify and locate regional faults.

Some other variation techniques used in this domain that derivative from classification is shown in Table 1. E.g., for neural network-based approach, the prediction errors of a certain number of the latest time steps that above a threshold will indicate a collective anomaly. Image segmentation-based approaches will divide the graph into groups of highly related elements and calculate a score for each element indicating how different it is from the rest of its group. The classifier-based approach usually uses the sliding window technique to train a classifier to differentiate normal patterns and abnormal patterns.

3.2 Time Series Dataset

For time series dataset, the temporal continuity is much stronger than symbolic dataset. The dataset can be modeled as a matrix with n columns and an unbounded number of rows, each data observation in time series dataset can be described as: a_i=(<b_{1i},b_{2i}, …,b_{ni}>, t_i), t is a temporal attribute, b is a behavioral attribute, n represents the number of features that selected to describe the target behavioral attribute, and i records the tag of the collected order of each data. The behavioral attribute has a strong correlation with its temporal attribute, both of them must be involved in the detection model. Collective anomalies in time series are often defined as abnormalities in temporal continuity (Mikalsen et al., 2018).

As shown in Figure 9, the collective anomaly in such datasets is represented by abnormal behavioral patterns under certain temporal constraints. The purpose of detection is to find a subsequence that is relatively abnormal to the rest of the entire dataset within a certain time interval. According to whether there are criteria to judge abnormity, the paper divides the existing detection approaches into the following three types.

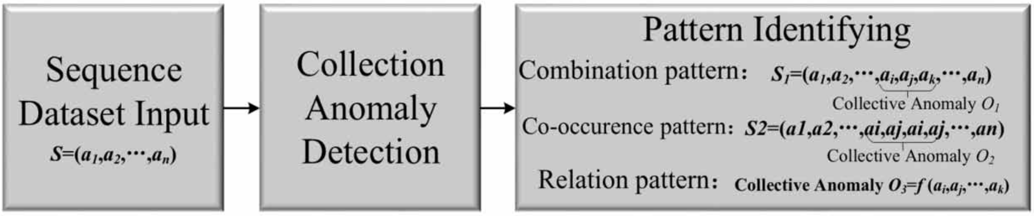
3.2.1 Detection Approach Based on Similarity Measurement

In such applications, if the criteria for determination is clear while the normal data patterns in the dataset are unknown or partially known, the collective anomaly existing in the dataset can be attributed to the similarity anomaly pattern.

Table 1. Representative classification techniques used for collective anomaly detection

Technique used	Research Focus	Classical References
Techniques based on neural network	Fitting and forecasting	(Bontemps, Cao, Mcdermott, & Le-Khac, 2016) (Prado-Romero & Gago-Alonso, 2017)
Techniques based on image segmentation	Feature extraction and segmentation rules	(Jia et al., 2018) (Araya, Grolinger, Elyamany, Capretz, & Bitsuamlak, 2016)
Techniques based on classifier training	Prior knowledge and training	(Bian, Hui, Sun, Zhao, & Tan, 2019)
Techniques based on machine learning	Sample labeling and transfer learning	(Dai, Cao, Wang, Deng, & Yang, 2019)

Figure 9. Collective anomaly detection model of time series dataset


Sequence Dataset Input
S=(a1,a2,⋯,an)
Collection Anomaly Detection
Pattern Identifying
Combination pattern: S1=(a1,a2,⋯,ai,aj,ak,⋯,an)
Collective Anomaly O1
Co-occurence pattern: S2=(a1,a2,⋯,ai,aj,ai,aj,⋯,an)
Collective Anomaly O2
Relation pattern: Collective Anomaly O3=f (ai,aj,⋯,ak)

In this case, if normal patterns are attainable and there is enough labeled data to represent these patterns, collective anomaly detection methods rely on the similarity measure functions which are constructed by selecting appropriate features according to the abnormal judgment criteria. The anomaly score of a test sequence is defined as the anomalous degree that compares to normal sequences.

Such as in the fields of particle physics and molecular chemistry, the result of each reaction is a sequence dataset which will be produced by a specific mechanism. The results of known mechanisms can be treated as normal patterns, and the observations that make up these results will be bound to follow certain distributions(Uddin & Choudhury, 2019). The similarity measure function used for anomaly detection can be constructed according to the statistical features of the data distribution. The anomaly score of unknown sequences (which may be the result of a faulty process or an undiscovered mechanism) can be measured by the distributional difference from normal patterns, and the data contained in the abnormal sequences will be considered as collective anomaly.

For example, Marteau (2020) built a fixed-background model with density estimation: p_{FB}=(1-λ) p_B+λp_A, λ is the proportionality coefficient. According to the distribution of normal sample sequences reflected by $pB(x)$ and the distribution of the entire dataset reflected by $pFB(x)$, it intends to detect collective anomalies with abnormal distribution $pA(x)$. Maru and Kobayashi (2020) chose the regression model-based approach, Weng and Liu (2019) selected Gaussian model-based approaches, and H. Qin, X. Zhan, and Y. Zheng (2021) selected mixture distribution based approaches by analyzing data distribution, the degree of abnormality can be measured by either an abnormal feedback in mean, variance, or both.

3.2.2 Detection Approach Based on Probabilistic Measurement

In this case, there is no criterion to determine whether a pattern is normal or abnormal, all patterns will be normal under certain conditions, the detection model is built to find a set of normal transition probabilities between different patterns. If the transition probability from the former pattern to the current pattern does not reach the threshold, the current pattern will be regarded as an abnormal pattern, and sequences that constitute the pattern will be treated as collective anomaly.

The most widely used detection approach to solve such problems is the Markov-based model, whose operation process can be divided into two phases: training and testing (Pang et al., 2019). The normal transfer probabilities of each pattern are calculated by training on the normal sample dataset, and the measured threshold of abnormity will be set accordingly. Then we tested the practical datasets to detect anomaly patterns with a transfer probability below the threshold.

Yisroel et al. (2017) calculated normal transfer probabilities between different device operations based on the *pc-Stream* algorithm (Mirsky, Shapira, Rokach, & Elovici, 2015) and output them as a sequence represented by a first-order Markov chain: $p=(\ldots p_{t-1},p_t)$. If the transformational probabilities between different device operations do not reach the threshold, it will be identified as an abnormal operation (which may be caused by theft), and the sequences that make up the abnormal operation will be regarded as collective anomaly. Some other kinds of Markov-based detection approach, for example, fuzzy Markov-based technique (X. Li, Gao, & Chen, 2021), sparse Markov based techniques (Ide, Khandelwal, & Kalagnanam, 2017) are all variations of the conventional Markov technique.

3.2.3 Detection Approach Based on Expected Trend Measurement

In this case, not only is there no criterion to judge whether the pattern is normal or abnormal, but the content of each pattern is hard to define. The most commonly used detection approach for this kind of datasets is the segmentation-based model (Trauer, Pfingstl, Finsterer, & Zimmermann, 2021). The method is that it first divides the dataset into segments and extracts the eigenvalue of each segment to represent the characteristics of the data within the segment. Then, the problem of collective anomaly detection will be converted to the problem of how to detect anomalies in the sequence composed of these eigenvalues. If an eigenvalue is significantly abnormal from the expected trend, its corresponding segment will be regarded as an anomaly, and the observations in this section will be treated as a collective anomaly. For such methods, the key to ensure its effectiveness lies in that the selection of data set segmentation points must be supported by sufficient domain knowledge.

The deficiency of such approaches is that the collective anomaly mined from the dataset often have no practical meaning. Although a lot of research has been devoted to this area, it has not been completely solved, which will be the focus of future work.

4. DETECTION ON SPATIAL SEQUENCE DATASET

The collective anomaly detection of spatial sequence dataset focuses on searching such subsequences, which may be normal in a single dataset, but will be abnormal when multiple datasets are analyzed together (Y. Li, Wang, Liu, Xian, & Xie, 2018). Therefore, an appropriate detection method must refer to the correlation analysis of multiple datasets, and detection techniques must consider both temporal and spatial attributes.

As mentioned above, spatial sequence datasets can be subdivided into homogeneous datasets and heterogeneous datasets. As shown in Figure 10, the detection method for homogeneous datasets focuses on the data fusion between multiple datasets. The detection method for heterogeneous datasets focuses on the joint analysis of various characteristics that are used to describe attributes within multiple datasets (N. Li, Chang, & Liu, 2020).

In this section, based on this classification, the paper first proposes the collective anomaly pattern that may appear on each of them. Then, for each pattern, the corresponding detection method is designed, and typical techniques for each method are enumerated.

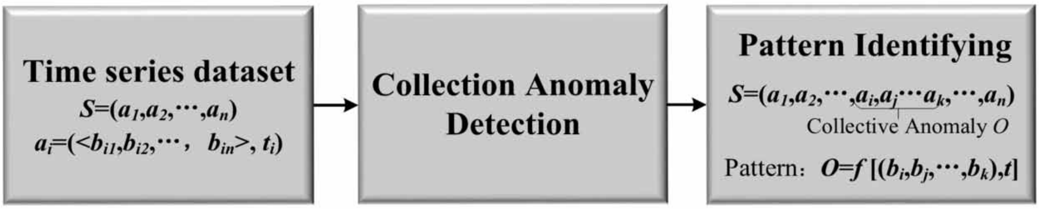
4.1 Homogeneous Datasets

In this section, based on this classification, the paper first proposes the collective anomaly pattern that may appear on each of them. Then, for each pattern, the corresponding detection method is designed, and typical techniques for each method are enumerated.

As mentioned above, homogeneous datasets are used to reflect the target problem by describing similar behavioral attributes, these datasets enjoy similar data distribution and structure (Xu, Jiang, Zeng, & Li, 2015). For example, in Fig 5, different kinds of traffic datasets, e.g., bike renting dataset, taxi flow dataset, and social media dataset, which are collected from different regions to describe urban dynamics and rhythms. These datasets generated in different regions of the city are used to reflect the same problem (urban traffic) by describing various types of traffic during sequential time intervals. The generated location is the spatial attribute, the traffic flow information is the behavioral attribute, and the time interval of collection is the temporal attribute. By fitting data curves, these datasets obey a similar distribution, which completely conforms to the characteristics of homogeneous datasets.

According to the model mentioned above, Zheng *et al.* (2015) designed a detection approach consisting of three parts as follows. Firstly, multiple-source latent topic model, which is used for fitting the data distribution of each dataset, especially for those sparse ones. Then, the spatio-temporal likelihood ratio test model, which is used to fuse the anomaly degree of each dataset. Lastly, an

Figure 10. Collective anomaly detection model of spatial sequence dataset


Time series dataset
$S=(a_1,a_2,\cdots,a_n)$
$a_i=(<b_{i1},b_{i2},\cdots, b_{in}>, t_i)$
Collection Anomaly Detection
Pattern Identifying
$S=(a_1,a_2,\cdots,\underbrace{a_i,a_j\cdots a_k}_{},\cdots,a_n)$
Collective Anomaly O
Pattern： $O=f\,[(b_i,b_j,\cdots,b_k),t]$

anomaly candidate set generation algorithm, which is used to detect collective anomaly by diagnosing abnormal urban traffic events.

In the same way, multi-node datasets that are generated from sensor networks also consist of homogeneous datasets. In comparison to the single in-node test, Bosman *et al*. (2017) proposed the anomaly detection approach based on fusing information coming from aggregate neighborhood nodes can identify more potential events that are reflected by the collective anomaly. They also found that the effectiveness of multi-datasets fusion is related to the diffusive property of the datasets (spatio-temporal correlations), as well as various types of sensors, and network topological features.

Through sorting out of this kind of research, it can be found that the key point of collective anomaly detection of homogeneous datasets lies in the fusion of multiple datasets. Efficiency can be improved by mutual reinforcement of datasets. Thus, detection approaches that suit for dataset fusion should have two parts. Firstly, a technique is needed to estimate the data distribution for each homogeneous dataset, especially for those sparse datasets which are hard to measure separately. Secondly, for these datasets with different data types and scales, it is necessary to build a suitable model to integrate multiple datasets. After fusion, the collective anomaly detection of the new aggregate dataset can be handled according to the detection process of temporal sequence datasets in the previous section. Some fusion techniques used in this domain are shown in Table 2.

4.2 Heterogeneous Datasets

For heterogeneous datasets, each of them reflects the description of incompletely identical behavioral attributes but which are correlational (Yang, Liu, Li, Wang, & Yang, 2016). Each dataset may differ in data distribution, type, structure, and the granularity of the corresponding temporal attribute.

Multi-datasets from wireless sensor networks are used for collecting a large amount of information reflecting environmental aspects such as light, temperature, humidity, etc. Although these factors are reflected in different behavioral attributes, there is a certain correlation between them. They reflect the state of the environment together, and the unexpected change of environment states will be diagnosed as a collective anomaly. Hence, these datasets form heterogeneous datasets together, different generating sources can be considered as spatial attributes.

Due to the fact that these datasets are collected at different time intervals and the data characteristics are also different from each other, Dey *et al*. (2019) designed the detection approach based on coupling edge data analysis with cloud data analysis. The former exploits a fully unsupervised artificial neural network algorithm, and the cloud data analysis exploits a multiple parameterized edit distance algorithm.

For such sequence datasets, collective anomalies can only be represented by multiple behavior attributes described by multiple datasets. Even if the normal patterns and exception patterns of each dataset can be identified and expressed, it is impossible to judge whether a collective anomaly

Table 2. Representative fusion techniques used for homogeneous datasets

Technique Used	Research Focus	Classical References
Techniques based on evidence theory	Semantic understanding and fusion	(Xiao, 2018) (Chang, Wu, Liu, Yan, & Qu, 2019)
Techniques based on deep learning	Sample labeling and transfer learning	(Chen & Jahanshahi, 2018) (Kiran, Dilip, & Ranjith, 2018)
Techniques based on fusion weight	Weight design and expert knowledge	(He, Wang, Liu, & Luo, 2019)
Techniques based on statistic model	Parameter fitting and multivariate function design	(Kene & Choudhury, 2019)
Techniques based on feature optimized	Feature selection and optimization methods	(Ke, Wu, Wu, & Xiong, 2018)

has occurred or is about to occur by relying on only one dataset. The critical point of detection in heterogeneous datasets lies in the association analysis among different parameters of multiple datasets (Zhou, Ha, Hu, & Ma, 2021).

Approaches that adapt to it should consist of two parts. Firstly, for the specific problem, appropriate features which can describe the target behavioral attribute should be selected. Then, a model should be established to fit the correlational relation among these features. After the model is established, it can be detected according to the detection mode of the temporal sequence datasets mentioned in the previous section. Some examples of association analysis techniques used in this domain are shown in Table 3.

5. CONCLUSION AND FUTURE WORK

The research on collective anomaly detection has reached a certain height in many fields, and a large number of detection techniques have been proposed. In order to improve the researching pertinence of collective anomaly detection to find the essence and commonality of various problems, this paper aims to make a comprehensive and structured review of the existing research on collective anomaly detection of sequence datasets. Firstly, aim to make the detection approach more pertinent and flexible, the sequence dataset is divided into two categories, each of which contains two subcategories. Then, for each subcategory, collective anomaly patterns which may appear on it and according detection models are discussed respectively. Lastly, for each model, a representative detection technique and its several deformations are exhibited.

Such kind of research is still in its infancy, and further explorations are urgently needed in multiple directions, some insights on the future work are as follows:

1. Collective anomaly detection techniques for different types of sequence datasets are not easy to generalize to each other, but the detection models they rely on may have similarities within the framework of higher-level design. Although there are distinct differences between the detection techniques for handling these two types of time series datasets, they all depend on the detection model, which includes temporal and behavioral attributes. Future research should continue in depth from the perspective of data attributes and relations, the intrinsic patterns of collective anomalies of similar categories should be concluded to build some relatively general detection models. When dealing with a targeted problem, based on these models, the detection techniques should be developed on specific data characteristics.
2. Currently, the concept of collective anomaly is often related to the problems that used to describe specific practical applications. Thus, commonly used detection techniques may be difficult to meet the accuracy requirements, while niche targeting techniques are often highly targeted and

Table 3. Representative association analysis techniques for heterogeneous datasets

Technique Used	Research Focus	Classical References
Techniques based on probability method	Data fitting for multivariate distributions	(Ali & Angelov, 2018) (L. Wang & Liang, 2019)
Techniques based on fuzzy logic theory	Mathematical logic and uncertain reasoning	(Dl, Mw, & Zx, 2019)
Techniques based on swarm intelligence algorithm	Parameter Settings, constraints and convergence criteria	(Tang, Gu, Yang, & Fu, 2019)
Techniques based on clustering approach	Clustering criteria and iteration rules	(J. Wang, Gao, Liu, Sangaiah, & Kim, 2019) (Mojahed & Beatriz, 2017)

hard to generalize. Future research should focus on exploring data patterns that can reflect the practical meanings of similar applications and storing them as knowledge bases. The efficiency of collective anomaly detection in targeted applications can be improved by merging the knowledge representation into the design of detection technique.

3. In the current research status, most of the models and techniques applied in collective anomaly detection are offline due to their heavy dependence on sample datasets. Such detection mechanisms, which belong to feedback control, can hardly be applied in fields with real time and predictive requirements. Due to the features of collective anomaly that appeared in various kinds of sequence datasets, future research should focus on how to design online detection approaches by incorporating machine learning and deep learning theory continuously.

FUNDING

This research received no external funding.

ACKNOWLEDGMENT

This paper is supported by Ministry of Education Foundation of China under Grant No.21YJC630044, 21YJC790152. University Talent Program under Grant No.20RC53, No.2022AH051774

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Ahmed, M., & Mahmood, A. N. (2014). Network Traffic Pattern Analysis Using Improved Information Theoretic Co-clustering Based Collective Anomaly Detection. *International Conference on Security and Privacy in Communication Networks*.

Ahmed, M., & Mahmood, A. N. (2015). Novel Approach for Network Traffic Pattern Analysis using Clustering-based Collective Anomaly Detection. *Annals of Data Science*, *2*(1), 111–130. doi:10.1007/s40745-015-0035-y

Al-Ghaili, A. M., Ibrahim, Z. A., Hairi, S., Rahim, F. A., & Kasim, H. (2021). A Review of anomaly detection techniques in advanced metering infrastructure. *Bulletin of Electrical Engineering and Informatics*, *10*(1), 266–273. doi:10.11591/eei.v10i1.2026

Ali, A. M., & Angelov, P. (2018). *Anomalous behaviour detection using heterogeneous data*. Academic Press.

Araya, D., Grolinger, K., Elyamany, H. F., Capretz, M., & Bitsuamlak, G. (2016). *Collective contextual anomaly detection framework for smart buildings.* Paper presented at the International Joint Conference on Neural Networks. doi:10.1109/IJCNN.2016.7727242

Auth, H. D. M. (1980). *Identification of Outliers*: Identification of Outliers.

Bian, J., Hui, X., Sun, S., Zhao, X., & Tan, M. (2019). A Novel and Efficient CVAE-GAN-Based Approach With Informative Manifold for Semi-Supervised Anomaly Detection. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 88903–88916. doi:10.1109/ACCESS.2019.2920251

Bigdeli, E., Raahemi, B., Mohammadi, M., & Matwin, S. (2015). *A fast noise resilient anomaly detection using GMM-based collective labelling.* Paper presented at the Science & Information Conference. doi:10.1109/SAI.2015.7237166

Bontemps, L., Cao, V. L., Mcdermott, J., & Le-Khac, N. A. (2016). *Collective Anomaly Detection Based on Long Short-Term Memory Recurrent Neural Networks*. Springer International Publishing. doi:10.1007/978-3-319-48057-2_9

Bosman, H. H., Iacca, G., Tejada, A., Wörtche, H. J., & Liotta, A. (2017). Spatial anomaly detection in sensor networks using neighborhood information. *Information Fusion*, *33*(C), 41–56. doi:10.1016/j.inffus.2016.04.007

Cappozzo, A., Greselin, F., & Murphy, T. B. (2020). Anomaly and Novelty detection for robust semi-supervised learning. *Statistics and Computing*, *30*(1), 1545–1571. doi:10.1007/s11222-020-09959-1

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, •••, 3.

Chang, X., Wu, J., Liu, H., Yan, X., & Qu, Y. (2019). Travel mode choice: A data fusion model using machine learning methods and evidence from travel diary survey data. *Transportmetrica*, *15*(2), 1587–1612.

Chen, F. C., & Jahanshahi, R. (2018). NB-CNN: Deep Learning-based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion. *IEEE Transactions on Industrial Electronics*, *65*(99), 4392–4400. doi:10.1109/TIE.2017.2764844

Chouhan, P., Richhariya, V., Chouhan, P., & Richhariya, V. (2015). A Survey: Analysis of Current Approaches in Anomaly Detection. *Foundation of Computer Science*, *111*(17), 14–20.

Chowdhury, D. (2019). Supervised Machine Learning and Heuristic Algorithms for Outlier Detection in Irregular Spatiotemporal Datasets. Journal of Environmental Informatics. *Journal of Environmental Informatics*, *33*(1), 1–16.

Dai, H., Cao, J., Wang, T., Deng, M., & Yang, Z. (2019). Multilayer one-class extreme learning machine. *Neural Networks*, *115*, 11–22. doi:10.1016/j.neunet.2019.03.004 PMID:30921561

Dasgupta, S., Yoshizumi, T., & Osogami, T. (2016). *Regularized Dynamic Boltzmann Machine with Delay Pruning for Unsupervised Learning of Temporal Sequences*. IEEE. doi:10.1109/ICPR.2016.7899800

Deng, Q., Wang, K., Zhao, M., Zou, Z., & Chen, L. (2021). *Personalized Bundle Recommendation in Online Games*. Academic Press.

Dey, S., Ye, Q., & Sampalli, S. (2019). A Machine Learning Based Intrusion Detection Scheme for Data Fusion in Mobile Clouds Involving Heterogeneous Client Networks. *Information Fusion*, *49*, 205–215. doi:10.1016/j.inffus.2019.01.002

Ding, F., Wang, J., Ge, J., & Li, W. (2018). *Anomaly detection in large-scale trajectories using hybrid grid-based hierarchical clustering*. Academic Press.

Dl, A., Mw, A., & Zx, B. (2019). Heterogeneous multi-attribute nonadditivity fusion for behavioral three-way decisions in interval type-2 fuzzy environment. *Information Sciences*, *496*, 242–263. doi:10.1016/j.ins.2019.05.044

Feroze, A., Daud, A., Amjad, T., & Hayat, M. K. (2021). *Group Anomaly Detection: Past Notions*. Present Insights, and Future Prospects.

Ghosh, S., Li, J., Ca, O. L., & Ramamohanarao, K. (2017). Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. *Journal of Biomedical Informatics*, *66*, 19–31. doi:10.1016/j.jbi.2016.12.010 PMID:28011233

Gokcesu, K., & Kozat, S. S. (2017). Online Anomaly Detection With Minimax Optimal Density Estimation in Nonstationary Environments. *IEEE Transactions on Signal Processing*.

Golmohammadi, K., & Zaiane, O. R. (2017). *Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation.* Paper presented at the Springer, Cham. doi:10.1007/978-3-319-64283-3_24

Gomathi, K., & Umagandhi, R. (2016). An efficient fuzzy based anomaly detection using collective clustering algorithm. *Kongunadu Research Journal*, *3*(1), 81–83. doi:10.26524/krj135

He, X., Wang, T., Liu, W., & Luo, T. (2019). Measurement Data Fusion Based on Optimized Weighted Least-Squares Algorithm for Multi-Target Tracking. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 13901–13916. doi:10.1109/ACCESS.2019.2894641

Huang, X., Ren, M., & Hu, Z. (2020). An Improvement of K-Medoids Clustering Algorithm Based on Fixed Point Iteration. *International Journal of Data Warehousing and Mining*, *16*(4), 84–94. doi:10.4018/IJDWM.2020100105

Ide, T., Khandelwal, A., & Kalagnanam, J. (2017). *Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection.* Paper presented at the IEEE International Conference on Data Mining.

Jia, , Lei, , & Jun, & Zhuang. (2018). Visual Analysis of Collective Anomalies Using Faceted High-Order Correlation Graphs. *IEEE Transactions on Visualization and Computer Graphics*. PMID:30582546

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence Classification for Credit-Card Fraud Detection. *Expert Systems with Applications*, *100*(JUN), 234–245. doi:10.1016/j.eswa.2018.01.037

Ke, W., Wu, C., Wu, Y., & Xiong, N. N. (2018). A New Filter Feature Selection Based on Criteria Fusion for Gene Microarray Data. IEEE Access, 1-1.

Kene, A. P., & Choudhury, S. K. (2019). Analytical modeling of tool health monitoring system using multiple sensor data fusion approach in hard machining. *Measurement*, *145*, 118–129. doi:10.1016/j.measurement.2019.05.062

Kiran, B., Dilip, T., & Ranjith, P. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, *4*(2), 36. doi:10.3390/jimaging4020036

Langarica, S., Ruffelmacher, C., & Nunez, F. (2019). An Industrial Internet Application for Real-Time Fault Diagnosis in Industrial Motors. *IEEE Transactions on Automation Science and Engineering, PP*, (99), 1–12.

Li, N., Chang, F., & Liu, C. (2020). Spatial-temporal Cascade Autoencoder for Video Anomaly Detection in Crowded Scenes. *IEEE Transactions on Multimedia, PP*, (99), 1–1.

Li, X., Gao, T., & Chen, W. (2021). *Fuzzy dynamic Markov model for Time Series Anomaly Detection.* Paper presented at the 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). doi:10.1109/IPEC51340.2021.9421340

Li, Y., Wang, J., Liu, X., Xian, N., & Xie, C. (2018). *DIM Moving Target Detection using Spatio-Temporal Anomaly Detection for Hyperspectral Image Sequences.* Paper presented at the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. doi:10.1109/IGARSS.2018.8517601

Liang, X., Ali, M. Z., & Zhang, H. (2020). Induction Motors Fault Diagnosis Using Finite Element Method: A Review. *IEEE Transactions on Industry Applications*, *56*(2), 1205–1217. doi:10.1109/TIA.2019.2958908

Liao, W., Hou, D., & Jiang, W. (2019). An Approach for a Spatial Data Attribute Similarity Measure Based on Granular Computing Closeness. *Applied Sciences (Basel, Switzerland)*, *9*(13), 2628. doi:10.3390/app9132628

Lin, L., & Su, J. (2019). Anomaly detection method for sensor network data streams based on sliding window sampling and optimized clustering. *Safety Science*, *118*(10), 70–75. doi:10.1016/j.ssci.2019.04.047

Liu, Q., Klucik, R., Chen, C., Grant, G., Gallaher, D., Lv, Q., & Shang, L. (2017). Unsupervised detection of contextual anomaly in remotely sensed data. *Remote Sensing of Environment*.

Liu, Y., Guo, B., Li, N., Zhang, J., & Yao, L. (2019). DeepStore: An Interaction-Aware Wide&Deep Model for Store Site Recommendation With Attentional Spatial Embeddings. *IEEE Internet of Things Journal*, *6*(4), 7319–7333. doi:10.1109/JIOT.2019.2916143

Marquez-Grajales, A., Acosta-Mesa, H. G., & Mezura-Montes, E. (2017). *An adaptive symbolic discretization scheme for the classification of temporal datasets using NSGA-II.* Paper presented at the 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). doi:10.1109/ROPEC.2017.8261674

Marteau, P. F. (2020). *Random Partitioning Forest for Point-Wise and Collective Anomaly Detection -- Application to Intrusion Detection*. Academic Press.

Maru, C., & Kobayashi, I. (2020). *Collective Anomaly Detection for Multivariate Data using Generative Adversarial Networks.* Paper presented at the 2020 International Conference on Computational Science and Computational Intelligence (CSCI). doi:10.1109/CSCI51800.2020.00106

Matsushita, Y., Egami, K., Sawada, A., Saito, M., & Yoshida, S. (2019). Analyses of soil bacterial community diversity in naturally and conventionally farmed apple orchards using 16S rRNA gene sequencing. *Applied Soil Ecology*, *141*, 26–29. doi:10.1016/j.apsoil.2019.04.010

Mikalsen, K., Oyvind, B., & Filippo, M. (2018). Time series cluster kernel for learning similarities between multivariate time series with missing data. Pattern Recognition the Journal of the Pattern Recognition Society.

Mirsky, Y., Shapira, B., Rokach, L., & Elovici, Y. (2015). *pcStream: A Stream Clustering Algorithm for Dynamically Detecting and Managing Temporal Contexts.* Paper presented at the Springer International Publishing.

Mohiuddin & Ahmed. (2017). Thwarting DoS Attacks: A Framework for Detection based on Collective Anomalies and Clustering. *Computer, 50*(9), 76-82.

Mojahed, A., & Beatriz, D. L. I. (2017). An adaptive version of k -medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach. *Knowledge and Information Systems*, *50*(1), 27–52. doi:10.1007/s10115-016-0930-3

Pang, J., Liu, D., Peng, Yu., & Peng, X. (2019). Collective Anomalies Detection for Sensing Series of Spacecraft Telemetry with the Fusion of Probability Prediction and Markov Chain Model. *Sensors (Basel)*, *19*(3), 722. doi:10.3390/s19030722 PMID:30754619

Prado-Romero, M. A., & Gago-Alonso, A. (2017). *Detecting Contextual Collective Anomalies at a Glance.* Paper presented at the 23rd International Conference on Pattern Recognition (ICPR).

Prasanta, G., Bhattacharyya, D. K., Borah, B., & Kalita, J. K. (2011). A Survey of Outlier Detection Methods in Network Anomaly Identification. *The Computer Journal*, (4), 570–588.

Qin, H., Zhan, X., & Zheng, Y. (2021). *CSCAD: Correlation Structure-based Collective Anomaly Detection in Complex System*. Academic Press.

Shah, A., Azam, N., Ali, B., Khan, M. T., & Yao, J. T. (2021). A Three-way Clustering Approach for Novelty Detection. *Information Sciences*, *569*, 650–668. doi:10.1016/j.ins.2021.05.021

Tang, S., Gu, Z., Yang, Q., & Fu, S. (2019). *Smart Home IoT Anomaly Detection based on Ensemble Model Learning From Heterogeneous Data.* Paper presented at the 2019 IEEE International Conference on Big Data (Big Data). doi:10.1109/BigData47090.2019.9006249

Trauer, J., Pfingstl, S., Finsterer, M., & Zimmermann, M. (2021). Improving Production Efficiency with a Digital Twin Based on Anomaly Detection. *Sustainability (Basel)*, *13*(18), 13. doi:10.3390/su131810155

Trivedi, I., Monik, M., & Mridushi, M. (2016). Credit Card Fraud Detection. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(1), 39–42. doi:10.17148/IJARCCE.2016.5109

Truong, T., Hai, D., Le, B., Fournier-Viger, P., & Fujita, H. (2021). Efficient Algorithms for Mining Frequent High Utility Sequences with Constraints. *Information Sciences*, *568*(5), 239–264. doi:10.1016/j.ins.2021.01.060

Uddin, S., & Choudhury, N. (2019). Actor-Level Dynamicity: Its Distribution Analysis Eases Anomaly Detection in Longitudinal Networks. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 69422–69433. doi:10.1109/ACCESS.2019.2917256

Valencia, J. A., Astray, G., Fernández-González, M., Aira, M. J., & Rodríguez-Rajo, F. J. (2019). Assessment of neural networks and time series analysis to forecast airborne Parietaria pollen presence in the Atlantic coastal regions. *International Journal of Biometeorology*, *63*(1), 735–745. doi:10.1007/s00484-019-01688-z PMID:30778684

Verenich, I., Dumas, M., Rosa, M. L., Maggi, F. M., & Francescomarino, C. D. (2015). *Complex Symbolic Sequence Clustering and Multiple Classifiers for Predictive Process Monitoring*. Paper presented at the International Conference on Business Process Management.

Wang, J., Gao, Y., Liu, W., Sangaiah, A., & Kim, H. J. (2019). An Improved Routing Schema with Special Clustering Using PSO Algorithm for Heterogeneous Wireless Sensor Network. *Sensors (Basel)*, *19*(3), 671. doi:10.3390/s19030671 PMID:30736392

Wang, L., & Liang, Q. (2019). Representation Learning and Nature Encoded Fusion for Heterogeneous Sensor Networks. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 39227–39235. doi:10.1109/ACCESS.2019.2907256

Wen, H., Sun, J., Zeng, Q., Zhang, X., & Yuan, Q. (2018). The Effects of Traffic Composition on Freeway Crash Frequency by Injury Severity: A Bayesian Multivariate Spatial Modeling Approach. *Journal of Advanced Transportation*, *2018*(PT.4), 1–7. doi:10.1155/2018/6964828

Weng, Y., & Liu, L. (2019). A Collective Anomaly Detection Approach for Multidimensional Streams in Mobile Service Security. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 49157–49168. doi:10.1109/ACCESS.2019.2909750

Xiao, F. (2018). Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46.

Xiaodi, H., Minglun, R., & Zhongfeng, H. (2020). A novel approach to identify regional fault of urban power grid based on collective anomaly detection. *International Journal of Modelling Identification and Control*, *36*(1), 78–87. doi:10.1504/IJMIC.2020.115389

Xin, W., & Xu, Y. (2011). *Detecting anomalies in symbolic sequence dataset*. IEEE.

Xu, J., Jiang, Y., Zeng, C., & Li, T. (2015). Node anomaly detection for homogeneous distributed environments. *Expert Systems with Applications*, *42*(20), 7012–7025. doi:10.1016/j.eswa.2015.04.037

Yang, S., Liu, Z., Li, J., Wang, S., & Yang, F. (2016). Anomaly Detection for Internet of Vehicles: A Trust Management Scheme with Affinity Propagation. *Mobile Information Systems, 2016*(1), 1-10.

Yin, C., Sun, Z., Jin, W., & Kim, J. U. (2015). *An Improved K-Means Using in Anomaly Detection.* Paper presented at the 2015 First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA). doi:10.1109/CCITSA.2015.11

Yin, C., Zhang, S., & Kwang-Jun, K. (2017). Mobile Anomaly Detection Based on Improved Self-Organizing Maps. *Mobile Information Systems* , 1-9.

Yisroel, , Asaf, , & Bracha, & Rokach. (2017). Anomaly detection for smartphone data streams. *Pervasive and Mobile Computing*.

Yuan, Y., Xiong, Z., & Wang, Q. (2019). VSSA-NET: Vertical Spatial Sequence Attention Network for Traffic Sign Detection. *IEEE Transactions on Image Processing*, *28*(7), 1–1. doi:10.1109/TIP.2019.2896952 PMID:30716035

Zamini, M., & Hasheminejad, S. (2019). A comprehensive survey of anomaly detection in banking, wireless sensor networks, social networks, and healthcare. *Intelligent Decision Technologies*, *13*(2), 1–42. doi:10.3233/IDT-170155

Zhang, K., Wang, B., & Kuo, C. (2022). PEDENet: Image anomaly localization via patch embedding and density estimation. *Pattern Recognition Letters*, *153*, 153. doi:10.1016/j.patrec.2021.11.030

Zhang, R., Xia, H., Shao, S. S., Ren, H., & Cheng, X. G. (2020). An Intrusion Detection Scheme Based on Repeated Game in Smart Home. *Mobile Information Systems*, *2020*(1), 1–9. doi:10.1155/2020/8844116

Zheng, Y., Zhang, H., & Yu, Y. (2015). Detecting collective anomalies from multiple spatio-temporal datasets across different domains. *Sigspatial International Conference on Advances in Geographic Information Systems*. doi:10.1145/2820783.2820813

Zhifeng, Z., Meng, L., & Rongpeng, L. (2017). Temporal-spatial distribution nature of traffic and base stations in cellular networks. *IET Communications*, *11*(16), 2410–2416. doi:10.1049/iet-com.2017.0330

Zhou, Z., Ha, M., Hu, H., & Ma, H. (2021). Half Open Multi-Depot Heterogeneous Vehicle Routing Problem for Hazardous Materials Transportation. *Sustainability (Basel)*, *13*(3), 13. doi:10.3390/su13031262

Xiaodi Huang received the B.S. degree from Anhui Science and Technology University in 2011, the M.S. degree from Yunnan Minzu University in 2014, and the Ph.D. degree from Hefei University of Technology in 2020. He is currently a Lecturer with Hefei University. His main research interests include data mining and fault diagnosis.

Po Yun received the M.S. degrees in Anhui University in 2016, and the Ph.D. degree from Hefei University of Technology in 2017. He is currently a Lecturer with Hefei University. His main research interests include nerual network technique and fault diagnosis.

Zhongfeng Hu received the B.S. degree from Guizhou University in 2008, the M.S. degree from Guilin University of Technology in 2012, and the Ph.D. degree from Hefei University of Technology in 2017. He is currently a Lecturer with Hefei University. His main research interests include intelligent manufacturing.