Expert Systems with Applications 42 (2015) 1998–2012

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Minimal infrequent pattern based approach for mining outliers in data streams

C. Sweetlin Hemalatha *, V. Vaidehi, R. Lakshmi

Department of Information Technology, Madras Institute of Technology, Anna University, Chennai, India

ARTICLE INFO

Article history:
Available online 13 October 2014

Keywords:
Minimal infrequent pattern
Outlier detection
Data streams
Data mining

ABSTRACT

Outlier detection is an important task in data mining which aims at detecting patterns that are unusual in a dataset. Though several techniques are proved to be useful in solving some outlier detection problems, there are certain issues yet to be resolved. Most of the existing methods compute distance of points in full dimensional space to detect outliers. But in high dimensional space, the concept of proximity may not be qualitatively meaningful due to the curse of dimensionality and incurs high computational cost. Moreover, the existing methods focus on discovering outliers but do not provide the interpretability of different subspaces that cause the abnormality. Frequent pattern mining based approaches resolve the aforementioned issues. Recently, infrequent pattern mining has attracted the attention of data mining research community which aims at discovering rare associations and researches in this area motivated to propose a new method to detect outliers in data streams. Infrequent patterns are more interesting than frequent patterns in some domains such as fraudulent credit transactions, anomaly detection, etc. In such applications, mining infrequent patterns facilitates detecting outliers. Minimal infrequent patterns are generators of family of infrequent patterns. In this paper, a novel method is presented to detect outliers by mining minimal infrequent patterns from data streams. Three measures namely Transaction Weighting Factor (TWF), Minimal Infrequent Deviation Factor (MIPDF) and Minimal Infrequent Pattern based Outlier Factor (MIFPOF) are defined. An algorithm called Minimal Infrequent Pattern based Outlier Detection (MIFPOD) method is proposed for detecting outliers in data streams based on mined minimal infrequent patterns. The effectiveness of the proposed method is demonstrated on synthetic dataset obtained from vital dataset collected from body sensors and a publicly available real dataset. The experimental results have shown that the proposed method outperforms the existing methods in detecting outliers.

1. Introduction

Infrequent pattern mining in data streams deals with extracting rare or unusual patterns from stream of data. Initially, pattern mining was focused on discovering frequent patterns (Agrawal, Imielinski, & Swami, 1993; Agrawal, Mannila, Srikant, Toivonen, & Verkamo 1996), i.e., those patterns whose frequency of occurrence (support) exceeds predefined minimum threshold. Recently, mining infrequent but useful patterns has gained the attention of data mining research community. In contrast to static dataset, mining useful patterns in data streams is a challenging task as it consists of an unbounded sequence of data which arrive to the system in a continuous manner. Windowing techniques such as sliding window (Chang & Lee, 2004; Chi, Wang, Yu, & Muntz, 2006; Lee, Lin, & Chen, 2005), damped window (Chang & Lee, 2003) and landmark window (Li, Lee, & Shan, 2004; Manku & Motwani 2002; Yu, Chong, Lu, Zhang, & Zhou, 2006) have been proposed to capture the characteristics of evolving data streams. Mining frequent patterns finds application in market basket analysis, click stream analysis, web link analysis, genome analysis, etc. However, mining infrequent patterns are more important and relevant compared to frequent patterns in certain applications like network intrusion detection, credit card fraud detection and anomaly detection.

An outlier is an observation or a point that varies considerably from other data (Hawkins, 1980). Detection of outliers by mining infrequent patterns can definitely provide a promising solution for finding threats. Minimal infrequent patterns (Haglin & Manning, 2007) act as seeds for generating entire family of infrequent patterns. In this paper, a method for detecting outliers in data streams by extracting minimal infrequent patterns is presented.

* Corresponding author.
E-mail addresses: sweetlinh@gmail.com (C. Sweetlin Hemalatha), vaidehi@annauniv.edu (V. Vaidehi), lakshmi.ravi90@yahoo.com (R. Lakshmi).

http://dx.doi.org/10.1016/j.eswa.2014.09.053
0957-4174/

The mined minimal infrequent patterns are used as descriptors of outliers. The basic idea is that those observations that form the source for generation of minimal infrequent patterns are likely to be an outlier as it possesses "discriminating feature values" compared to other observations under consideration. New measures such as Transaction Weighting Factor (TWF), Minimal Infrequent Pattern Deviation Factor (MIPDF) and Minimal Infrequent Pattern based Outlier Factor (MIFPOF) are defined for detecting outliers and Minimal Infrequent Pattern based Outlier Detection (MIFPOD) algorithm is proposed in this paper.

The main contribution of this paper is summarized as follows:

- An algorithm for mining Minimal Infrequent Patterns from Data Streams (MIP-DS) is presented.
- Three simple measures such as TWF, MIPDF and MIFPOF are defined for outlier detection.
- MIFPOD algorithm for detecting outliers based on minimal infrequent patterns is proposed.

The rest of the paper is organized as follows. Section 2 presents related work in the area of pattern mining and outlier detection. Section 3 introduces some preliminaries of pattern mining that are relevant to this paper. In Section 4, an example and an algorithm for mining minimal infrequent patterns are given. The definition of new outlier factors based on minimal infrequent patterns, an example and a novel algorithm for outlier detection are also presented. Experimental results are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Related work

Most of the existing pattern mining algorithms focus on mining frequent patterns. However, mining infrequent patterns has also been meaningful rather than frequent patterns in certain scenarios.

2.1. Infrequent pattern mining

Liu, Hsu, and Ma (1999) proposed MSapriori algorithm in order to address the problem of mining infrequent patterns that provide high confidence rules. The algorithm employs multiple minimum support thresholds to each item in the database. The minimum support of the rule is defined in terms of minimum support of the items that appear in the rule. Thus, the user is allowed to specify different support thresholds for different rules. However, the change in threshold values is driven by a subjective parameter, making the algorithm sensitive to user preferences.

Koh and Rountree (2005) developed Apriori-Inverse algorithm that defines both minimum and maximum support thresholds for generating rare items and discarding extremely rare items. The algorithm searches in level-wise bottom up fashion similar to Apriori algorithm. During each iteration, only those patterns whose support lies between minimum and maximum support are considered for further processing. The algorithm captures only the perfect rare items and fails to capture itemsets that are infrequent but frequent as individual items.

Adda, Wu, and Feng (2007) proposed an algorithm called AfRIM for mining infrequent patterns. Unlike Apriori, the algorithm begins to search in top down fashion for identifying rare patterns as they may not occur at the top of the lattice. Besides rare items, this algorithm also explores patterns with zero support, which may be inefficient.

Troiano and Scibelli (2014) proposed a time efficient algorithm called Rarity that uses top-down approach similar to AfRIM. Though the proposed algorithm explores all non-zero rare patterns in less time, it demands high memory requirements.

All the above algorithms adopt Apriori approach directly or indirectly which will generate candidates and hence involve pruning steps. Thus, much time is spent for searching the non-rare items besides finding rare items.

Ashish, Akshay, and Arnab (2011) proposed an algorithm based on pattern growth paradigm to find infrequent patterns. The algorithm constructs header table for each item sets which is linked to the pattern growth tree containing all item transaction. The authors have used two more structures namely projected tree and residual tree. Projected tree is constructed by removing the items which were frequent and residual tree is constructed to reduce space.

Tsang, Koh, and Dobbie (2011) proposed a frequent pattern (FP) tree based algorithm for generating a set of rare items. According to the algorithm, entire database is scanned only once to find infrequent patterns whose support is less than minimum support. Authors have used information gain component to identify a set of rare association rules.

The aforementioned algorithms are based on mining infrequent patterns in static datasets. But mining infrequent patterns in streaming data needs modifications to suite the characteristics of data streams. Huang, Koh, and Dobbin (2012) proposed a new algorithm called SRP (Streaming Rare Pattern tree) to generate a set of rare items. In this tree based approach, for each incoming transaction the items are inserted into tree based on FP growth approach. Generally, FP tree is constructed after arranging all items in the transactions in descending order of the support. But in the case of data stream arranging the items altogether is not possible. To overcome this problem the authors maintained a structure called connection table which keeps track of items in the window in canonical order. If an item has support less than minimum support then the path containing the item generates the all subset of infrequent items.

The above mentioned methods provide solution for discovering either all rare patterns or only non-zero rare patterns. Szathmary, Napoli, and Valtchev (2007) presented an algorithm called Minimal Rare Generator (MRG) to find both rare as well as frequent itemset. Authors have used three parameters namely predefined support, occurrence of each item and a key. Each itemset is given a predefined support and a key with a value 'yes' if the items' predefined support and evaluated support are equal and 'no' otherwise. Only the itemset with key value 'yes' are considered for next iteration. Thus both rare as well as frequent item sets are encountered at each step. Haglin and Manning (2007) designed an algorithm called MINIT (Minimally Infrequent Item set) for mining minimal infrequent items. The algorithm works by sorting and ranking items based on the support. At each step, the items having support less than minimum support are chosen and only the transaction containing those items are selected for further processing.

Both MRG and MINIT are based on mining infrequent patterns from static dataset. This paper proposes a multi-pass non-overlapping sliding window based algorithm for mining Minimal Infrequent Patterns in Data Streams (MIP-DS).

2.2. Outlier detection methods

A lot of techniques for detecting outliers have been proposed in the last several years (Chandola, Banerjee, & Kumar, 2009; Pimentel, Clifton, Clifton, & Tarassenko, 2014). The statistics community has performed extensive research on outlier detection (Barnett & Lewis, 1994). Existing approaches to outlier detection can be classified into seven categories: (1) distribution based methods (2) distance based methods (3) density based methods (4) clustering based methods (5) Artificial Neural Network based methods (6) information theoretic based methods (7) frequent pattern mining based methods.

2.2.1. Distribution based methods

A standard distribution model such as Normal, Gamma, etc. is used to fit the dataset. Those objects that deviate from the postulated model are recognized as outliers. Yamanishi, Takeuchi, and Williams (2000) modeled the normal behaviors using Gaussian mixture model and scored each datum based on the changes in the model. This method has been combined with supervised learning to obtain a general pattern for outliers (Yamanishi & Takeuchi, 2001). Aggarwal and Philip (2008) used probability density function (pdf) to model uncertain data and detect outliers by applying a microcluster definition. Hido, Tsuboi, Kashima, Sugiyama, and Kanamori (2011) proposed a density ratio estimation based method to detect outliers. The authors have used direct density ratio estimation method called unconstrained least squares importance fitting (uLSIF) that allows optimization of tuning parameters without subjective trial and error. The major disadvantage of these methods is that the underlying distribution of data is unknown in practice and assumes a Gaussian distribution for the training data and hence may not give satisfactory results.

2.2.2. Distance based methods

Knorr and Ng (1998) and Knorr, Ng, and Tucakov (2000) have proposed distance based outlier method which tags an object X as an outlier if at least a fraction p of objects in a dataset D having a distance greater than $d$min from X. Though this method is both intuitive and computationally feasible, it depends on two subjective parameters p and $d$min. To overcome the shortcomings of distance based outlier method, Ramaswamy, Rastogi, and Shim (2000) further extended it based on the distance of a point from its kth neighbor. All the points are ranked by distance to its kth neighbor and top-k points are identified as outliers. Angiulli and Pizzuti (2002) proposed an alternative algorithm in which the outlier factor of a data point is evaluated as the sum of distances from its k-nearest neighbors. These methods generalize the notion of distribution based methods and enjoy better computational complexity for larger k values. A novel distance measure named centered mahalanobis distance is introduced to detect outliers (Todeschini, Ballabio, Consonni, Sahigara, & Filzmoser, 2013). It differs from the classical method by centering the covariance matrix at each data sample rather than at the data centroid.

Yu, Wang, Xiao, and Yang (2010) proposed a distance based detection on uncertain data that focus on evaluating the uncertainty in the existence of a sensor at a particular location based on a confidence score. Shaikh and kitagawa (2012) extended the idea to address the uncertainty that lies in the sensed data using bounded Gaussian distribution. However, existing models describe uncertainty of discrete data. Yan, Xia, and Feng (2012) reported abnormal pattern detection in continuous uncertain data by re-expressing Euclidean distance according to probability density function and got a probabilistic metric to compute the dissimilarity of two uncertain series. Distance based outlier detection technique for data streams proposed by Anguiulli and Fassetti (2007) uses sliding window to form a bounded transaction dataset and thus detect outliers inside the window. Ishida and Kitagawa (2008) and Kontaki, Gounaris, Papadopoulos, Tsichlas, and Manolopoulos (2011) have also addressed the challenges in detecting outliers over data streams. However, distance based methods need many subjective parameters that require trial and error to attain desired results. Also, the time complexity of these methods increases with increase in dimension of data.

2.2.3. Density based methods

Breunig, Kriegel, Ng, and Sander (2000) introduced the notion of density based local outliers and proposed a measure called Local Outlier Factor (LOF) of objects that specifies the objects' degree of being an outlier. LOF of an object is obtained by considering the clustering structure in the bounded rectangle of the object. Jiang, Tseng, and Su (2001) regarded small clusters as outliers by evaluating each point in small clusters. He, Xu, and Deng (2003) introduced Cluster Based Local Outlier Factor (CBLOF) that measures both the size of the cluster to which the object belongs and the distance between the object and its nearest cluster. Kriegel and Zimek (2008) proposed angle based outlier detection for high dimensional data that outperforms LOF. A new outlier factor named Uncertain Local Outlier Factor (ULOF) was proposed to handle uncertain dataset (Liu & Deng, 2013). Liu, Xiao, Cao, Hao, and Deng (2013) proposed a hybrid approach for detecting outliers by generalizing the Support Vector Data Description (SVDD) in order to cope with data uncertainty. The authors have introduced a confidence value which is associated with each data besides the class label and which measures the strength of the corresponding label. The major drawbacks of these methods are that they suffer curse of dimensionality, consume more time and space for processing and hence not suitable for large data.

2.2.4. Clustering based methods

Outliers are obtained as by-products of clustering (Jain, Murty, & Flynn, 1999; Jiang & An, 2008). Duan, Xu, Liu, and Lee (2009) presented a new algorithm that identifies small clusters as outliers by evaluating it as a whole rather than each point in a small cluster. Elahi, Li, Nisar, Lv, and Wang (2008) presented an efficient outlier detection method for data streams based on K-means clustering. The proposed method maintains the candidate outliers and mean of every cluster to decide outliers. Ren, Wu, Zhang, and Hu (2009) extended that approach to heterogeneous data streams. Shao, Böhm, Yang, and Plant (2010) proposed a method that detects outliers based on clustering by synchronization (Böhm, Plant, Shao, & Yang, 2010) that detects clusters of arbitrary shape, size and density using Minimum Description Length (MDL) principle. However, these methods are optimized for finding clusters and not for detecting outliers.

2.2.5. Artificial Neural Network based methods

Neural Network based methods generalizes data knowledge and enables continuous learning. The unsupervised and adaptive neural networks such as Self organizing Maps (SOM), Adaptive Resonant Theory (ART) and Grow When Required (GWR) have been employed to address the problem of outlier detection. Marsland, Shapiro, and Nehmzow (2002) developed a neural network that can grow when required by inserting new nodes when the node that best matches the input could not be found. Albertini and De Mello (2007) proposed a network that combines the best features of ART, SOM and GWR and dynamically creates a new neuron when no existing neurons could classify the input pattern. Thus, creation of new neurons signals the presence of outliers. Barreto and Aguayo (2009) have showed that temporal SOM-based algorithms are suitable for identifying anomalous patterns in time series rather than static competitive neural networks. However, these methods require different threshold parameters to be set experimentally. Kit, Sullivan, and Ballard (2011) have used growing neural gas network (Fritzke, 1995), an unsupervised incremental clustering algorithm to detect changes in the input distribution over time. García-Rodríguez et al. (2012) have shown that self organizing neural network models are capable of satisfying temporal constraints in managing real-time applications by inserting multiple neurons per iteration where the number of such neurons is controlled dynamically, However, using large number of neurons complicates the process of obtaining a representation of the distribution of training data with good accuracy in real-time.

2.2.6. Information theoretic based methods

Jiang, Sui, and Cao (2010) have proposed a novel definition and algorithm for detecting outliers in rough sets based on Information

Entropy (IE). Filippone and Sanguinetti (2010) proposed an approach that estimates the expected information content of a data point to detect changes online. Li, Li, Wang, and Zhai (2014) presented incremental entropy based integrated framework for clustering categorical data streams. The authors have proposed Minimal Dissimilarity Data Labeling (MDDL) technique to assign a proper cluster label for an incoming data point in the current sliding window based on the clustering result of the previous window. Those data points that cannot be exactly marked are considered as outliers. Wu and Wang (2013) proposed an optimization model for detecting outliers in large scale categorical data using a new concept called holoentropy that takes both entropy and total correlation into consideration. But the time complexity of the Information theoretic based outlier detection algorithms increases with increase in dimension of data.

2.2.7. Frequent pattern based methods

Wei, Qian, Zhou, Jin, and Yu (2003) introduced hypergraph model for mining outliers in categorical dataset. However, the method requires every data point to be represented as a vertex of a hypergraph and hence it is computationally intensive. He, Xu, Huang, and Deng (2005) presented a novel method for detecting outliers by mining frequent patterns from the dataset. A new measure called Frequent Pattern Outlier Factor (FPOF) has been defined which adds the support of frequent patterns in transactions to detect the outliers. A transaction t is said to be an outlier if it contains less frequent patterns and for which FPOF value will be low. The drawback of this method is that the support of a frequent pattern and its subsets are added resulting in duplicate computation and thus the measure FPOF cannot reflect the normal degree of transactions accurately. In order to overcome the drawback of FPOF based outlier method, Zhang, Wu, and Yu (2010) proposed an improved algorithm based on Longest Frequent Pattern (LFP algorithm). The authors have devised a measure called Longest Frequent Pattern Outlier Factor (LFPOF) which computes the ratio of length of the longest frequent pattern in a transaction to the length of the transaction. Zhou, Sun, Zhang, and Yang (2007) presented Weighted Frequent Pattern Outlier Factor (WFPOF) for detecting outliers in data streams. Wu and Ma (2011) proposed Weighted Closed Frequent Pattern Outlier Factor (WCFPOF) to detect outliers in categorical streaming data by mining closed frequent patterns in a sliding window. Lin, Le, and Bo (2010) focused on detecting outliers in high dimensional time series data stream based on Maximal Frequent Pattern Outlier Factor (MFPOF). Kao and Huang (2012) proposed a rule base outlier detection method that segments the transactions from data streams, mines approximate frequent patterns in a single scan and detects outliers based on association rules. However, these methods rely on mining frequent patterns (FP) or variants of frequent patterns followed by computing a measure based on respective patterns for detecting outliers. Hence, the computational time increases exponentially with the number of attributes.

In this paper, we propose a novel technique for detecting outliers that overcomes the aforementioned shortcomings of existing approaches and has the following features:

- Detects outliers in transaction data streams and hence feasible for instant identification of outliers.
- Handles data of any distribution.
- Considers the subspace that causes abnormality (outliers) rather than full dimensional space.
- Emphasizes on mining minimal infrequent patterns rather than frequent patterns and hence efficient in computational point of view.
- Introduces an outlier scoring measure based on discovered minimal infrequent patterns which will be more appropriate than frequent pattern based scoring method in the case of rare point/events detection.
- Detects both single and group outliers.
- Reports the possible patterns responsible for outlier transactions in data streams opposed to existing methods which just detect outliers.

3. Preliminaries

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items that represents a piece of information in a domain. A set $X = \{i_p, i_q, \ldots, i_r\} \subseteq I$ and $p, q, r \in [1, n]$ is called an itemset or a pattern and if $|X| = k$, then X is called as k-itemset. A transaction t_i is a tuple that contains set of items in I and is characterized by transaction identifier, *tid*.

A data stream $DS = [t_1, t_2, \ldots, t_N)$ can be formally defined as an infinite sequence of transactions and $N \to \infty$. Sliding window SW model allows processing of only the most recent transactions from the data stream DS and $|SW|$ defines the size of the sliding window i.e., the number of recent transactions considered for processing. The count of an itemset X in a sliding window SW, denoted by $Count_{SW}(X)$ is defined as number of transactions in SW that contain X. The support of a pattern X is defined as number of transactions in SW that contain X to total number of transactions in the sliding window SW as given in Eq. (1).

$$Support_{SW}(X) = \frac{Count_{SW}(X)}{|SW|} \tag{1}$$

Definition 1. A pattern X is said to be frequent if its support in the window SW exceeds the predefined minimum support threshold δ as given by Eq. (2).

$$Support_{SW}(X) \geqslant \delta \tag{2}$$

Definition 2. A pattern X is said to be infrequent or rare if its support in the window SW does not exceed the predefined minimum support threshold δ as given by Eq. (3).

$$Support_{SW}(X) < \delta \tag{3}$$

Definition 3. A pattern X is said to be a Minimal Infrequent Pattern (MIFP) if it is infrequent and all of its proper subsets are frequent as given by Eq. (4).

$$\text{MIFP} = \{x | support(x) < \delta \wedge (\nexists y | y \subset x \wedge support(y) < \delta)\} \tag{4}$$

Definition 4. (*Hawkins–Outlier*). An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

4. Outlier detection based on minimal infrequent patterns

This paper addresses the problem of detecting the presence of outliers by mining minimal infrequent patterns in data streams using non overlapping sliding window model. This section has two subsections. The first subsection presents an example for mining Minimal Infrequent Patterns in Data Streams (MIP-DS) followed by an algorithm to discover them. The next subsection presents some definitions concerning MIP based outliers in data streams, an example and algorithm to find Minimal Infrequent Pattern based Outlier Detection (MIFPOD).

4.1. An example – minimal infrequent pattern mining

The incoming transactions within the sliding window *SW* is represented in the form of binary matrix $M{:}R \times C \rightarrow \{0,1\}$ where *R* is the number of transactions and *C* is the number of items that characterize the domain.

Fig. 1 shows a sample transaction set and the corresponding lattice representing all possible patterns for sample dataset. The dashed circles represent infrequent patterns (*DE*, *ADE*, *BDE*, *ABDE* and *ABCDE*) and bold dashed circle represents the seed *DE* responsible for generation of other infrequent patterns and it is the minimal infrequent pattern. In the case of data streams, whole dataset is not known in advance and hence they must be processed as and when they come. In order to process the most recent data, a sliding window of *SW* of size 3 i.e., $|SW| = 3$ is considered. Tables 1 and 2 present the canonical ordering of items in the sliding window and representation of transactions in binary matrix form respectively. In Tables 1 and 2, Tid represents the transaction identifier $\langle T1, T2, T3\rangle$ which may be a timestamp in the case of data stream and *Items* represents the attribute list $\langle A,B,C,D,E\rangle$. The value '1' under each item represents presence of normal data and '0' represents presence of abnormal data. Let the minimum support threshold $\delta = 0.6$ i.e., the item should occur at least twice in a window of size 3 to be a frequent item.

Those patterns whose support does not exceed predefined minimum support threshold δ are infrequent patterns. According to Table 2, *C* is an infrequent pattern of length 1, as its support is less than δ. All frequent patterns of length $l > 1$ are processed in ascending order of their support as frequent patterns with less support have a good chance to become infrequent patterns. During each step, those patterns identified as infrequent are not considered in the subsequent steps as mining only minimal infrequent patterns are of interest. Hence, following *C*, the item *E* is analyzed for possible infrequent patterns associated with $E\backslash C$ as *C* has been processed in the previous step. The pattern *DE* of length 2 is infrequent whereas all other patterns *AE* and *BE* of length 2 are frequent. Following the item *E*, items $B\backslash C$ and *DE* and $A\backslash C$ and *DE* are processed and the only possible pattern is *AB* which is frequent. Therefore, only those patterns that are not processed in the previous step are considered for subsequent processing. Thus, the two minimal infrequent patterns mined from the sample shown in Table 2 are *C* and *DE*.

Table 1
Canonical ordering of items in $|SW| = 3$.

Tid	Items
*T*1	{*A*,*B*,*D*,*E*}
*T*2	{*A*,*B*,*D*}
*T*3	{*A*,*B*,*C*,*E*}

Table 2
Transaction in binary matrix form.

Tid/items	*A*	*B*	*C*	*D*	*E*
*T*1	1	1	0	1	1
*T*2	1	1	0	1	0
*T*3	1	1	1	0	1

Fig. 2 shows the multi pass Minimal Infrequent Pattern mining algorithm for data streams (MIP-DS). The algorithm takes a binary matrix *M*, sliding window size $|SW|$ and minimum support threshold δ as input and generates all possible minimal infrequent patterns. The set $\mathcal{CG}_1$ denotes patterns of length 1 whose support is calculated by Eq. (1) under Definition 1. The support values are sorted in ascending order so as to give high priority to the patterns with less support. Those patterns satisfying Eq. (4) under Definition 3 constitute the minimal infrequent patterns of length 1. For each frequent pattern, the method *genCandidate*(*i*) generates all possible patterns of length > 1 for the argument *i* and the same steps are repeated for mining all minimal infrequent patterns.

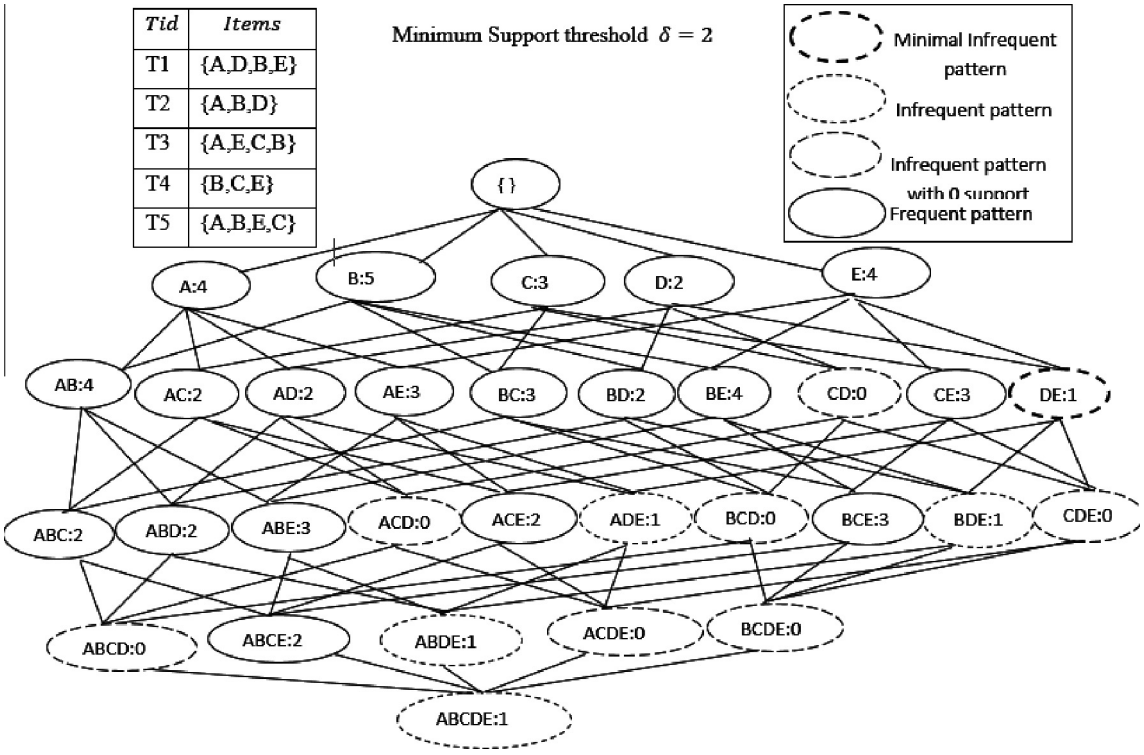

Fig. 1. A sample data set and the corresponding lattice showing all possible patterns.

MIP-DS algorithm

Input: Binary Matrix $M: R \times C \rightarrow \{0,1\}$, Sliding window size $|SW|$, Minimum support threshold δ

Output:Minimal Infrequent Patterns

1. $\mathcal{CG}_1 \leftarrow \{1-itemset\}$
2. $Support^j_{SW}(\mathcal{CG}_1) = \frac{Count_{SW}(X)}{|SW|}, j \in \mathcal{CG}_1$
3. $S \leftarrow Sort\left(Support^j_{SW}(\mathcal{CG}_1)\right)$
4. $mIF_1 \leftarrow \left\{ j \in \mathcal{CG}_1 \mid Support^j_{SW} < \delta \right\}$
5. For each item $i \in S \setminus mIF_1$ do
6. $\quad \mathcal{CG}_{next} \leftarrow genCandidate(i)$
7. $\quad$ Calculate $Support_{SW}(\mathcal{CG}_{next})$
8. $\quad mIF_{next} \leftarrow \{ Y \mid (Support_{SW}(\mathcal{CG}_{next}) < \delta) \wedge (\nexists y \subset Y | support(y) < \delta) \}$
9. End for

Fig. 2. Algorithm for mining minimal infrequent patterns from data streams.

4.2. Minimal infrequent pattern based outliers

Minimal Infrequent Pattern based Outlier Detection in data streams adopts Hawkins' definition for outlier as given by Definition 4. That is, given a sliding window *SW* of stream of data, any transaction $t \in SW$, if t has some properties that deviate much from other transactions in *SW*, then transaction t is identified as an outlier with respect to other stream of data in sliding window *SW*. Minimal infrequent patterns mined within *SW* are used as an effective measure for detecting outliers in *SW*.

Definition 5. Transaction Weighting Factor (TWF). Let $DS = [t_1, t_2, \ldots, t_N)$ be N transactions within sliding window *SW* and let MIFP be the set of all minimal infrequent patterns mined from the transactions in *SW*. For each transaction t, the Transaction Weighting Factor of t is defined as:

$$\mathrm{TWF}(t) = \frac{\sum_{x \subseteq t, x \in \mathrm{MIFP}} |x|}{|\mathrm{MIFP}|} \tag{5}$$

where MIFP is given by Eq. (4) in Definition 3.

If a transaction t contains unusual values, then it forms a source for generating one or more minimal infrequent patterns, MIFP. The TWF(t) value of such transactions will be less and it is more likely to be an outlier.

Definition 6. Minimal Infrequent Pattern Deviation Factor (MIPDF). Let $DS = [t_1, t_2, \ldots, t_N)$ be N transactions within sliding window *SW* and predefined minimum support threshold δ. For each minimal infrequent pattern a, the Minimal Infrequent Pattern Deviation Factor of a is defined as:

$$\mathrm{MIPDF}(a) = \left(\frac{\delta - support(a)}{|\mathrm{MIFP}|} \right) \tag{6}$$

If the support of a pattern does not exceed predefined minimum support threshold δ, then it finds its place in the set of minimal infrequent patterns, MIFP. But, how far it deviates from δ determines the degree of outlier. Hence, the contribution of MIPDF is also taken into account for detecting outliers.

Definition 7. Minimal Infrequent Pattern based Outlier Factor (MIFPOF). For each transaction t, the Minimal Infrequent Pattern based Outlier Factor is defined as:

$$\mathrm{MIFPOF}(t) = 1 - \left(\mathrm{TWF}(t)^* \left(\sum_{x \in \mathrm{MIFP} \wedge x \subseteq t} \mathrm{MIPDF}(x) \right) \right) \tag{7}$$

where TWF(t) is the Transaction Weighting Factor of t and MIPDF(x) is the Infrequent Pattern Deviation Factor of pattern x belonging to MIFP and a subset of transaction t.

4.2.1. An example

Consider the sample transaction set shown in Table 2 and let the outlier threshold v be 0.7. Fig. 3 shows the abnormal values marked with circles and associated patterns are C and DE which are the minimal infrequent patterns for the sample transaction set. Hence, minimal infrequent patterns can clearly capture possible outliers. Fig. 4 presents the Minimal Infrequent Pattern based Outlier Detection (MIFPOD) algorithm.

$$Count(A) = 3, \quad Count(B) = 3, \quad Count(C) = 1,$$
$$Count(D) = 2, \quad Count(E) = 2$$

$$Support(A) = \frac{3}{3} = 1, \quad Support(B) = \frac{3}{3} = 1, \quad Support(C) = \frac{1}{3} = 0.3$$

$$Support(D) = \frac{2}{3} = 0.6, \quad Support(E) = \frac{2}{3} = 0.6$$

From Definition 2, the set of infrequent patterns *IP* are:

$$IP = \{C:0.3, AC:0.3, BC:0.3, CE:0.3, ABC:0.3, BCE:0.3, ABCE:0.3, DE:0.3, BDE:0.3, ABDE:0.3\}$$

Table 3
Specification of various vital parameters.

Sensors	Description of measuring data	Normal range for adults	Sampling rate
Heart rate (HR)	Frequency of cardiac cycle	60–100 beats/min	56 ms/data
Breathing rate (BR)	Respiration rate	16–20	56 ms/data
Blood pressure (BP)	Force exerted by circulating blood on the walls of blood vessels	Systolic:100–119 mmHg Diastolic: 70–79 mmHg	1 s/data
ECG	Electrical activity of the heart	Frequency: 0.5–100 Hz Amplitude: 0.25–1 mV	4 ms/data

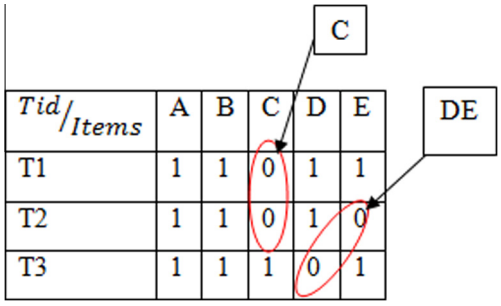

Fig. 3. Abnormal values and associated patterns.

MIFPOD Algorithm

Input: Data stream DS, Sliding window size $|SW|$, outlier threshold v

Output: set of transactions identified as outliers

1. Begin
2. Mine the set of Minimal Infrequent Patterns $MIFP$ using MIP-DS Algorithm
3. For each $x \in MIFP$ do
4. $MIPDF(x) = (\delta - support(x))/|MIFP|$
5. End for
6. For each transaction t in SW do
7. $T(t) = 0, A(t) = 0$
8. For each $x \in MIFP$ do
9. If t contains x then
   a. $T(t) = T(t) + 1$
   b. $A(t) = A(t) + MIPDF(x)$
10. End if
11. End for
12. $TWF(t) = T(t)/|SW|$
13. $\text{MIFP}OF(t) = 1 - (TWF(t) * A(t))$
14. If $(MIFPOF(t) > v)$, then output t.
15. End if
16. End for
17. End

Fig. 4. Algorithm for detecting *Minimal Infrequent Pattern based Outliers.*

From Definition 3, the set of minimal infrequent MIP are:

$\text{MIP} = \{C : 0.3, DE : 0.3\}$

From Definition 5,

$$\text{TWF}(T1) = \frac{1}{2} = 0.5, \quad \text{TWF}(T2) = \frac{0}{2} = 0, \quad \text{TWF}(T3) = \frac{1}{2} = 0.5$$

From Definition 6,

$$\text{MIPDF}_{T1}(C) = \left(\frac{0.6 - 0.3}{2}\right) = \frac{0.3}{2} = 0.15,$$

$$\text{MIPDF}_{T1}(DE) = \left(\frac{0.6 - 0.3}{2}\right) = \frac{0.3}{2} = 0.15$$

From Definition 7,

$$\text{MIPOF}(T1) = 1 - (0.5^{*}(0.15 + 0.15)) = 1 - 0.15 = 0.85 > v$$

$$\text{MIPOF}(T2) = 1 - 0 = 1 > v$$

$$\text{MIPOF}(T3) = 1 - (0.5^{*}(0.15 + 0.15)) = 1 - 0.15 = 0.85 > v$$

Therefore $T1$, $T2$ and $T3$ are MIP based outliers in the sliding window SW of data streams DS but with varied outlier factors.

Consider a scenario where item A in Table 2 happens then it does not and then it happens again.

Case 1: If the minimum support threshold is 0.6 (i.e., count/window size = 2/3), then item A is frequent as per the Definition 1.
Case 2: If the minimum support threshold is 0.3 (i.e., count/window size = 1/3), then item A is frequent as per the Definition 1.
Case 3: If the minimum support threshold is 3 (i.e., count/window size = 3/3), then item A is infrequent as per the Definition 2. Besides item A, items C, D and E are also infrequent. Hence, the minimum support threshold has effect on minimal infrequent patterns generated but not on the accuracy of MIFPOD algorithm for detecting outliers.

5. Experimental results

To evaluate the proposed outlier method, we have considered a synthetic dataset and a real dataset. The synthetic dataset is obtained by embedding abnormal patterns by randomly flipping the bits in the binary representation of corresponding vital parameters in the real data collected from a normal person using Biosensors that measures vital parameters. The real dataset used in the experiment is Wisconsin breast cancer data obtained from UCI Machine Learning Repository (Hettich & Bay, 1999) for detecting health anomaly (outlier). Initially, experiment is conducted on synthetic dataset to verify that the proposed method can find relationship between minimal infrequent patterns and outliers and then on the real dataset to prove the effectiveness of the proposed MIFPOD algorithm. The performance of the proposed algorithm against traditional distance based outlier (Knorr & Ng, 1998; Knorr et al., 2000), K-nearest neighbor algorithm (Ramaswamy et al., 2000) and FP-based outlier algorithm (He et al., 2005) is reported in this section.

5.1. Physiological data from Biosensors

Two Biosensor nodes namely Bioharness 3 (Zephyr Bioharness) and Blood Pressure sensor node (Zephyr Pressure Monitor) shown in Fig. 5 are used for physiological data acquisition. Bioharness 3 chest strap body sensor node measures Heart Rate (HR), Breathing Rate (BR) and ECG amplitude. Wrist worn Blood pressure sensor node measures systolic and diastolic Blood Pressure (BP). These parameters are termed as vital parameters or vital signs and they form the vital dataset.

Fig. 5 shows the block diagram for Minimal Infrequent Pattern based Outlier Detection (MIFPOD). The Biosensors form the Body Area Network (BAN) which communicates with Pattern Mining and Outlier Detection System via Bluetooth. A sliding window is defined to process the recent stream of health data. Minimal infrequent patterns are mined from the windowed data using MIP-DS algorithm and outliers are detected using the proposed MIFPOD algorithm which is based on minimal infrequent patterns.

The binary representation of the vital dataset is obtained by setting a bit value 1 to the normal range of values for vital parameters (Potter, Perry, Castaldi, Stockert, & Hall, 2011) as given by Table 3 and 0 otherwise. The last column of Table 3 specifies the sampling rate of vital parameters. As the vital dataset is collected from a

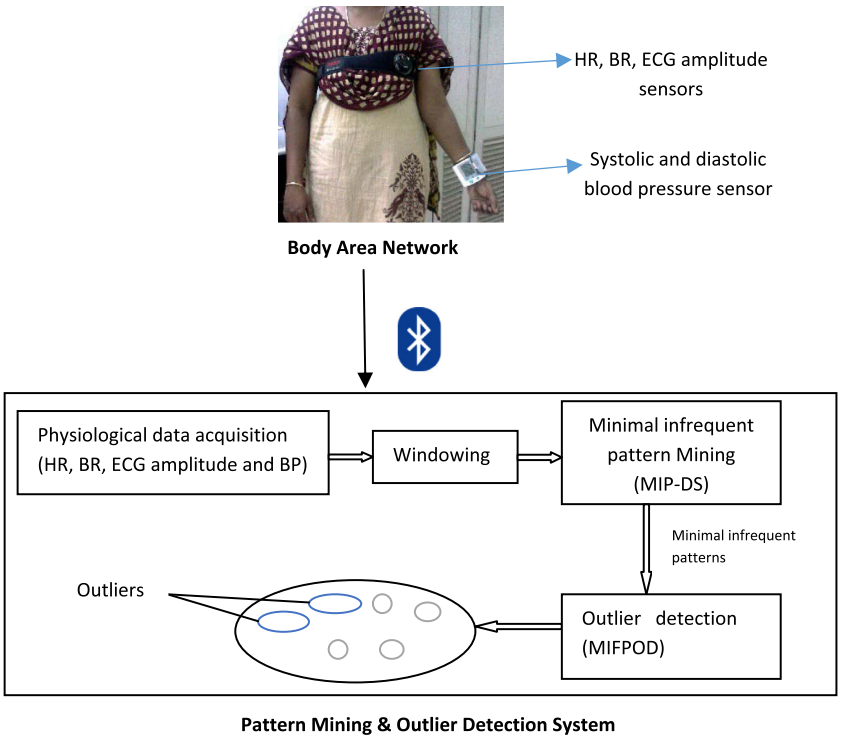

Fig. 5. Block diagram for Minimal Infrequent Pattern based Outlier Detection.

normal person, the bit pattern will be all 1s indicating normal health status with respect to the vital parameters considered. The synthetic data is then obtained by randomly flipping the bits of corresponding vital parameter to simulate abnormal health data. Such simulated abnormal data is embedded once between every five normal data. Any vital parameter (HR, BR, ECG amplitude and BP) having a value '0' indicates abnormal health status and the streams that contain such abnormal value are considered as an outlier. As outliers are detected in data streams, a sliding window is defined to process recent stream and minimum support threshold is defined to mine minimal infrequent patterns.

In Tables 4 and 8, the "WINDOW SIZE" denotes the size of the sliding window, and "MINIMUM SUPPORT" denotes the user specified minimum support threshold. In Tables 4, 5, 7 and 8, "NO. OF RARE CLASS" in each window specifies the number of transactions belonging to rare class, the "TOP RATIO (%) (NO. OF RECORDS)" denotes the percentage of transaction selected from the sliding window whose outlier factor calculated by the proposed MIFPOD method are higher than other transactions in the window. The "NO. OF RARE CLASS INCLUDED" is the number of transactions in the sliding window that belongs to abnormal category (rare class). The "COVERAGE RATIO" is the ratio of the "NO. OF RARE CLASS INCLUDED" to the number of transactions in the sliding window that belong to rare class.

The window size determines the number of data stream observations considered for mining minimal infrequent patterns. This parameter is subjective in nature and has been tested with varied sizes viz. 30, 60, 70 and 100. However, yet another subjective parameter, the "minimum support threshold" influence the mining of minimal infrequent patterns. It is set more close to window size as the health parameters are so sensitive and even small deviation from normal range is an outlier.

The proposed Minimal Infrequent Pattern based Outlier Detection (MIFPOD) method relies on discovering minimal infrequent patterns in the predefined window of observations. Hence, the parameter minimum support threshold decides the mined patterns. If it is set to a small value, then most patterns will be frequent and does not favor outlier detection. Hence, it is varied from 80% to 95% in order to mine the minimal infrequent patterns and both window size and minimum support threshold are presented in Tables 4 and 7 for vital dataset and Wisconsin breast cancer dataset respectively.

Table 4 summarizes the experimental results obtained for non-overlapping sliding window of varied sizes viz. 30, 60, 70 and 100 with outcomes of five successive windows under each size and varied minimum support thresholds. As abnormal data is simulated to occur once in every five transaction streams, the "NO. OF RARE CLASS" that exists in each window is "WINDOW SIZE"/5. The "NO. OF RECORDS" for top '*k*' outliers is varied randomly up to "NO. OF RARE CLASS" present in the respective window and TOP RATIO (%) is evaluated. The "NO. OF RARE CLASS INCLUDED" is equal to the "NO. OF RECORDS" as the detected top 'k' outliers are true outliers. The "COVERAGE RATIO" is 100% when "NO. OF RARE CLASS INCLUDED" is equal to "NO. OF RARE CLASS" present in the respective window. Since the successive windows in each size have the same number of rare classes, the outlier results reported are also the same. It is observed from the results that the proposed method is capable of reporting all the outliers present in the respective window with top ratio of 20%. Thus, minimal infrequent pattern mining that precedes outlier detection facilitates detection of true outliers faster.

The performance of the proposed MIFPOD algorithm (MIFP) is compared with FP-based outlier detection (FP), traditional distance based (DIS) and KNN based (KNN) outlier detection and the experimental results for five successive windows each of size 30 and minimum support threshold as 83% are summarized in Table 5. The comparison results show that the MIFP-based method outperforms other methods by capturing top 6 outliers with top ratio of 20% for vital dataset. It is because the mined minimal infrequent patterns are closely related to outliers than frequent patterns.

Table 4
Experimental results for vital dataset.

Window size = 30 minimum support = 25									
I window no. of rare class = 6		II window no. of rare class = 6		III window no. of rare class = 6		IV window no. of rare class = 6		V window no. of rare class = 6	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
13(4)	4(67)	13(4)	4(67)	13(4)	4(67)	13(4)	4(67)	13(4)	4(67)
20(6)	6(100)	20(6)	6(100)	20(6)	6(100)	20(6)	6(100)	20(6)	6(100)
Window size = 60 minimum support = 55									
I window no. of rare class = 12		II window no. of rare class = 12		III window no. of rare class = 12		IV window no. of rare class = 12		V window no. of rare class = 12	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
7(4)	4(33)	7(4)	4(33)	7(4)	4(33)	7(4)	4(33)	7(4)	4(33)
13(8)	8(67)	13(8)	8(67)	13(8)	8(67)	13(8)	8(67)	13(8)	8(67)
9(9)	9(75)	9(9)	9(75)	9(9)	9(75)	9(9)	9(75)	9(9)	9(75)
20(12)	12(100)	20(12)	12(100)	20(12)	12(100)	20(12)	12(100)	20(12)	12(100)
Window size = 70 minimum support = 65									
I window no. of rare class = 14		II window no. of rare class = 14		III window no. of rare class = 14		IV window no. of rare class = 14		V window no. of rare class = 14	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No of rare class included (coverage ratio)	Top ratio (%) no. of records	No of rare class included (coverage ratio)
6(4)	4(29)	6(4)	4(29)	6(4)	4(29)	6(4)	4(29)	6(4)	4(29)
11(8)	8(57)	11(8)	8(57)	11(8)	8(57)	11(8)	8(57)	11(8)	8(57)
13(9)	9(64)	13(9)	9(64)	13(9)	9(64)	13(9)	9(64)	13(9)	9(64)
17(12)	12(86)	17(12)	12(86)	17(12)	12(86)	17(12)	12(86)	17(12)	12(86)
20(14)	14(100)	20(14)	14(100)	20(14)	14(100)	20(14)	14(100)	20(14)	14(100)
Window size = 100 minimum support = 95									
I window no. of rare class = 20		II window no. of rare class = 20		III window no. of rare class = 20		IV window no. of rare class = 20		V window no. of rare class = 20	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
4(4)	4(20)	4(4)	4(20)	4(4)	4(20)	4(4)	4(20)	4(4)	4(20)
8(8)	8(40)	8(8)	8(40)	8(8)	8(40)	8(8)	8(40)	8(8)	8(40)
9(9)	9(20)	9(9)	9(20)	9(9)	9(20)	9(9)	9(20)	9(9)	9(20)
12(12)	12(60)	12(12)	12(60))	12(12)	12(60)	12(12)	12(60)	12(12)	12(60)
16(16)	16(80)	16(16)	16(80)	16(16)	16(80)	16(16)	16(80)	16(16)	16(80)
20(20)	20(100)	20(20)	20(100)	20(20)	20(100)	20(20)	20(100)	20(20)	20(100)

Table 5
Performance comparison of outlier detection methods with vital data.

MIFP		FP		DIS		KNN	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
Window 1 and no. of rare class = 6							
13(4)	4(67)	13(4)	0(0)	13(4)	1(17)	13(4)	1(17)
20(6)	6(100)	27(8)	1(17)	27(8)	1(17)	27(8)	1(17)
		30(9)	1(17)	30(9)	2(33)	30(9)	1(17)
		40(12)	2(33)	30(12)	2(33)	30(12)	1(17)
		53(16)	3(50)	40(16)	3(50)	40(16)	3(50)
		67(20)	4(67)	53(20)	3(50)	53(20)	4(67)
		80(24)	4(67)	80(24)	3(50)	80(24)	4(67)
		100(30)	6(100)	100(30)	6(100)	100(30)	6(100)
Window 2 and no. of rare class = 6							
13(4)	4(67)	13(4)	0(0)	13(4)	1(17)	13(4)	1(17)
20(6)	6(100)	27(8)	1(17)	27(8)	1(17)	27(8)	1(17)
		30(9)	1(17)	30(9)	2(33)	30(9)	1(17)
		40(12)	2(33)	30(12)	2(33)	40(12)	1(17)
		53(16)	3(50)	40(16)	3(50)	53(16)	3(50)
		67(20)	4(67)	53(20)	3(50)	67(20)	4(67)
		80(24)	4(67)	80(24)	5(83)	80(24)	4(67)
		100(30)	6(100)	90(27)	6(100)	100(30)	6(100)
Window 3 and no. of rare class = 6							
13(4)	4(67)	13(4)	0(0)	13(4)	1(17)	13(4)	3(50)
20(6)	6(100)	27(8)	1(17)	27(8)	2(33)	27(8)	3(50)
		30(9)	1(17)	30(9)	2(33)	30(9)	3(50)
		40(12)	2(33)	30(12)	3(50)	40(12)	3(50)
		53(16)	3(50)	40(16)	4(67)	53(16)	3(50)
		67(20)	4(67)	53(20)	5(83)	67(20)	3(50)
						100(30)	6(100)
Window 4 and no. of rare class = 6							
13(4)	4(67)	13(4)	0(0)	13(4)	1(17)	13(4)	0(0)
20(6)	6(100)	27(8)	1(17)	27(8)	3(50)	27(8)	0(0)
		30(9)	1(17)	30(9)	3(50)	30(9)	1(17)
		40(12)	2(33)	30(12)	3(50)	40(12)	2(33)
		53(16)	3(50)	40(16)	4(67)	53(16)	3(50)
		67(20)	4(67)	53(20)	5(83)	67(20)	6(100)
		80(24)	4(67)	80(24)	6(100)		
		100(30)	6(100)				
Window 5 and no. of rare class = 6							
13(4)	4(67)	13(4)	0(0)	13(4)	1(17)	13(4)	0(0)
20(6)	6(100)	27(8)	1(17)	27(8)	3(50)	27(8)	0(0)
		30(9)	1(17)	30(9)	3(50)	30(9)	1(17)
		40(12)	2(33)	30(12)	4(67)	40(12)	2(33)
		53(16)	3(50)	40(16)	4(67)	53(16)	2(33)
		67(20)	4(67)	53(20)	4(67)	67(20)	3(50)
		80(24)	4(67)	83(24)	5(83)	80(24)	3(50)
		100(30)	6(100)	100(30)	6(100)	90(27)	6(100)

Moreover, the proposed method also presents the subspace in the form of minimal infrequent patterns that cause abnormality (outliers) compared to full dimensional space in distance based and KNN based methods.

5.1.1. Limitations

The proposed method detects abnormal stream (outlier) by considering only vital parameters. There are some factors that can influence these parameters (Potter et al., 2011). For example, heart rate is influenced by factors such as physical exercise, posture (sitting, standing, lying), chronic heart disease, etc. The proposed method detects the data stream corresponding to the factors above as outliers but the person may not really be in abnormal health condition. Hence, the abnormality detection can be combined with human activity recognition for effective outlier detection.

5.2. Wisconsin breast cancer data

The Wisconsin breast cancer dataset contains 699 instances with 9 continuous attributes. Each record was categorized as benign or malignant. Benign is the commonly occurring class (normal) and malignant is the rare class (abnormal/outlier) as shown in Table 6. Data stream is considered as chunks of data such that every chunk contains fixed number of observations as specified by the size of the sliding window. Wisconsin breast cancer dataset is considered as a data stream by processing a set of samples as a chunk. Then, the sliding window is adjusted to process the next chunk. This strategy is followed till the entire dataset is processed.

Table 7 shows the experimental results of the proposed method for different window size $|SW|$ such as 30, 60 and 100 with varied minimum support threshold. The number of true outliers present in each window is specified by "NO. OF RARE CLASS". In each window size, the results obtained for five successive windows are reported. The top 'k' outliers detected by the proposed method

Table 6
Class distribution of Wisconsin breast cancer data.

Case	Class codes	Percentage of instances
Commonly occurring class	2	65.5%
Rare class	4	34.5%

Table 7
Experimental results for Wisconsin breast cancer dataset.

Window size = 30 minimum suport = 25									
I window no. of malignant = 9		II window no. of malignant = 20		III window no. of malignant = 13		IV window no. of malignant = 14		V window no. of malignant = 9	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	no .of rare class included (coverage ratio)	top ratio (%) no. of records	no .of rare class included (coverage ratio)
13(4)	4(44)	13(4)	4(20)	13(4)	4(31)	13(4)	4(29)	13(4)	4(44)
27(8)	8(89)	27(8)	8(40)	27(8)	8(62)	27(8)	8(57)	27(8)	8(89)
30(9)	9(100)	30(9)	9(45)	30(9)	9(69)	30(9)	9(64)	30(9)	9(100)
		40(12)	12(60)	40(12)	12(92)	40(12)	12(86)		
		53(16)	16(80)	43(13)	13(100)	47(14)	14(100)		
		67(20)	20(100)						
Window size = 60 minimum support = 55									
I window no. of malignant = 29		II window no. of malignant = 27		III window no. of malignant = 21		IV window no. of malignant = 31		V window no. of malignant = 29	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
7(4)	4(14)	7(4)	4(15)	7(4)	4(19)	7(4)	4(13)	7(4)	4(14)
13(8)	8(28)	13(8)	8(30)	13(8)	8(38)	13(8)	8(26)	13(8)	8(28)
15(9)	9(31)	15(9)	9(33)	15(9)	9(43)	20(12)	12(39)	15(9)	9(31)
20(12)	12(41)	20(12)	12(44)	20(12)	12(57)	27(16)	16(52)	20(12)	12(41)
27(16)	16(55)	27(16)	16(59)	27(16)	16(76)	33(20)	20(65)	27(16)	16(55)
33(20)	20(69)	33(20)	20(74)	33(20)	20(95)	40(24)	24(77)	33(20)	20(67)
40(24)	24(83)	40(24)	24(89)	40(24)	24(100)	47(28)	28(90)	40(24)	24(83)
47(28)	28(97)	45(27)	27(100)			52(31)	31(100)	48(29)	29(100)
48(29)	29(100)								
Window size = 100 minimum support = 95									
I window no. of malignant = 44		II window no. of malignant = 35		III window no. of malignant = 53		IV window no. of malignant = 34		V window no. of malignant = 34	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
4(4)	4(9)	4(4)	4(11)	4(4)	4(8)	4(4)	4(12)	4(4)	4(12)
8(8)	8(18)	8(8)	8(23)	8(8)	8(15)	8(8)	8(24)	8(8)	8(24)
9(9)	9(20)	9(9)	9(26)	9(9)	9(16)	9(9)	9(26)	9(9)	9(26)
12(12)	12(27)	12(12)	12(34)	12(12)	12(22)	12(12)	12(35)	12(12)	12(35)
16(16)	16(36)	16(16)	16(46)	16(16)	16(30)	16(16)	16(47)	16(16)	16(47)
20(20)	20(45)	29020)	20(57)	20(20)	20(38)	29(20)	20(59)	29(20)	20(59)
24(24)	24(55)	24(24)	24(69)	24(24)	24(45)	24(24)	24(71)	24(24)	24(71)
28(28)	28(64)	28(28)	28(80)	28(28)	28(53)	28(28)	28(82)	28(28)	28(82)
32(32)	32(73)	35(35)	35(100)	32(32)	32(60)	34(34)	34(100)	34(34)	34(100)
42(42)	42(95)			40(40)	40(75)				
44(44)	44(100)			44(44)	44(83)				
				53(53)	53(100)				

are true outliers and hence the NO. OF RARE CLASS INCLUDED" is same as that of "NO. OF RECORDS" and the "COVERAGE RATIO" is 100%. The value of TOP RATIO (%) differs as the number of outliers that exists in successive windows of same size varies. It is observed from the results that the proposed method successfully reports all true outliers present in the respective window. This is because the minimal infrequent patterns clearly characterize the outliers.

Table 8 shows the performance of the proposed MIFP based outlier detection method against FP-based outlier detection (FP), traditional distance based (DIS) and KNN based (KNN) outlier detection and the experimental results for five windows each of size 30 are summarized. For Wisconsin dataset, the performances of MIFP-based and FP-based methods are very close but better than distance based and KNN based methods.

Though the performances of MIFP-based and FP-based methods are same, it is observed from Fig. 6 that the time taken for detecting outliers based on minimal infrequent patterns is less compared to frequent patterns for varied number of transactions. It is because the process of generating minimal infrequent patterns from a sample of data takes lesser time than frequent patterns. Hence, considering infrequent patterns for anomaly (outlier) detection is more appropriate than frequent patterns and the proposed method can replace frequent pattern based outlier method and it suits applications that need faster response.

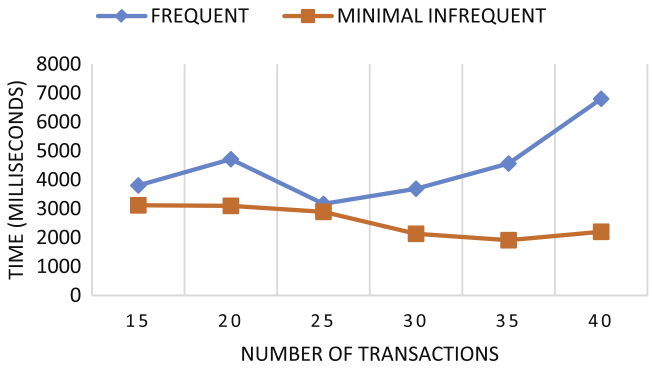

Fig. 6. Time taken for detecting outliers based on frequent vs. minimal infrequent patterns.

The effect of different parameters such as sliding window size and minimum support threshold on the output of the proposed algorithm is empirically studied. Fig. 7(a)–(d) depict the effect of parameter minimum support threshold on the output of the proposed algorithm for vital dataset processed in window size of 30, 60, 70 and 100 respectively. Similarly, Fig. 8(a)–(d) depict the effect of parameter minimum support threshold on the output of the proposed algorithm for Wisconsin breast cancer dataset processed in window size of 30, 60, 70 and 100 respectively. The minimum

Table 8
Performance comparison of outlier detection methods for Wisconsin breast cancer dataset.

MIFP		FP		DIS		KNN	
Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)	Top ratio (%) no. of records	No. of rare class included (coverage ratio)
Window 1 and no. of rare class = 9							
13(4)	4(44)	13(4)	4(44)	13(4)	3(33)	13(4)	3(33)
27(8)	8(89)	27(8)	8(89)	27(8)	5(56)	27(8)	5(56)
30(9)	9(100)	30(9)	9(100)	30(9)	5(56)	30(9) 40(12)	6(67) 9(100)
Window 2 and no. of rare class = 20							
13(4)	4(20)	13(4)	4(20)	13(4)	4(20)	13(4)	4(20)
27(8)	8(40)	27(8)	8(40)	27(8)	7(35)	27(8)	7(35)
30(9)	9(45)	30(9)	9(45)	30(9)	7(35)	30(9)	8(40)
40(12)	12(60)	40(12)	12(60)	40(12)	8(40)	40(12)	11(55)
53(16)	16(80)	53(16)	16(80)	53(16)	12(60)	53(16)	15(75)
67(20)	20(100)	67(20)	20(100)	67(20)	13(65)	67(20)	20(100)
				80(24)	15(75)		
				93(28)	17(85)		
				100(30)	20(100)		
Window 3 and no. of rare class = 13							
13(4)	4(31)	13(4)	4(31)	13(4)	4(31)	13(4)	4(31)
27(8)	8(62)	27(8)	8(62)	27(8)	8(62)	27(8)	4(31
30(9)	9(69)	30(9)	9(69)	30(9)	9(69)	30(9)	4(31)
40(12)	12(92)	40(12)	12(92)	40(12)	11(85)	40(12)	4(31
43(13)	13(100)	43(13)	13(100)	53(16)	11(85)	53(16)	5(38)
				67(20)	11(85)	67(20)	7(54)
				80(24)	11(85)	80(24)	8(62)
				100(30)	13(100)	93(28)	11(85)
						100(30)	13(100)
Window 4 and no. of rare class = 14							
13(4)	4(29)	13(4)	4(29)	13(4)	4(29)	13(4)	7(50)
27(8)	8(57)	27(8)	8(57)	27(8)	8(57)	27(8)	7(50)
30(9)	9(64)	30(9)	9(64)	30(9)	9(64)	30(9)	8(57)
40(12)	12(86)	40(12)	12(86)	40(12)	12(86)	40(12)	14(100)
47(14)	14(100)	47(14)	14(100)	47(14)	14(100)		
Window 5 and no. of rare class = 9							
13(4)	4(44)	13(4)	4(44)	13(4)	4(44)	13(4)	4(44)
27(8)	8(89)	27(8)	8(89)	27(8)	8(89)	27(8)	8(89)
30(9)	9(100)	30(9)	9(100)	30(9)	8(89)	30(9)	8(89)
				33(10)	9(100)	40(12)	8(89)
						53(16)	8(89)
						67(20)	8(89)
						80(24)	8(89)
						100(30)	9(100)

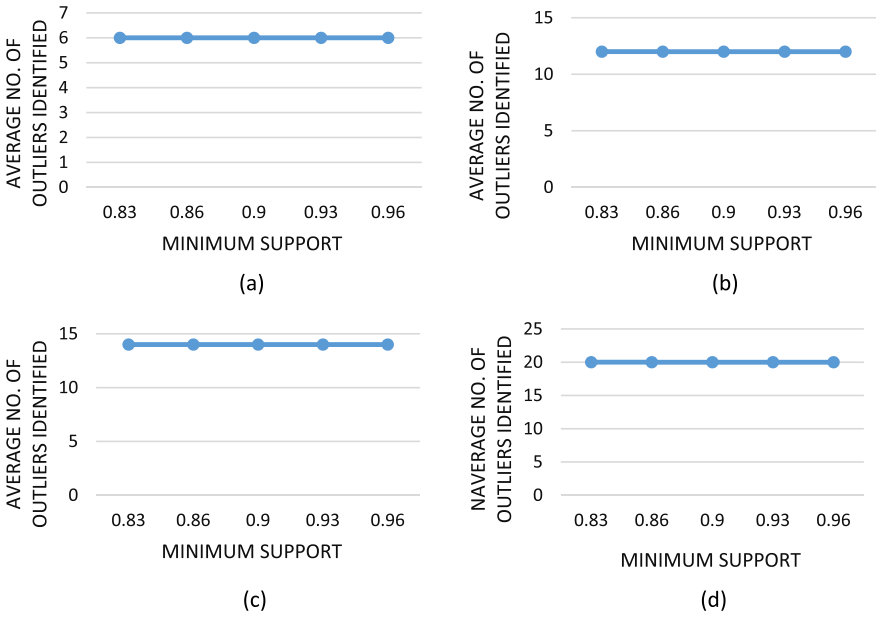

Fig. 7. Average number of outliers vs. different minimum support – vital dataset.

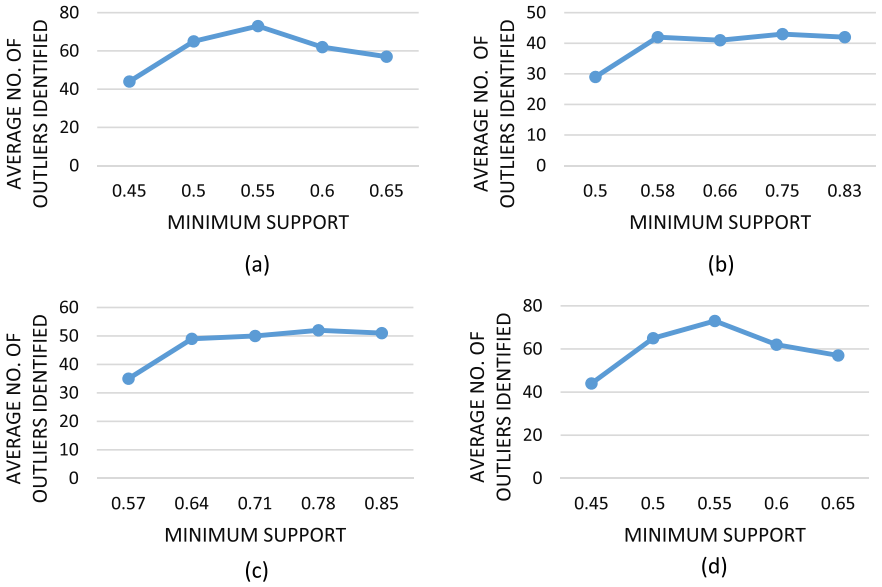

Fig. 8. Average number of outliers vs. different minimum support –Wisconsin breast cancer dataset.

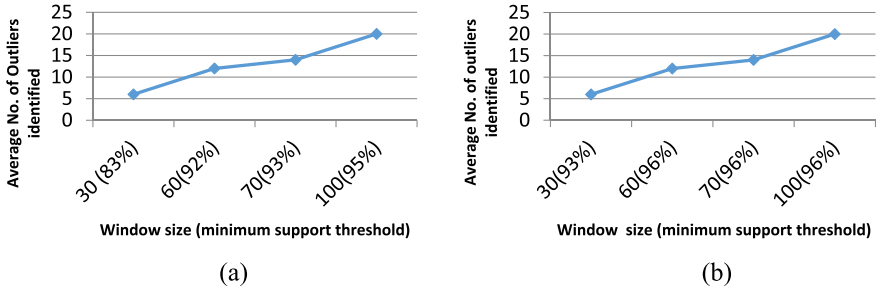

Fig. 9. Average number of outliers vs. Window size (minimum support threshold) – vital dataset.

support was varied and top ratio was fixed as 20%. It is observed that the user defined minimum support threshold does not affect the performance of the proposed outlier detection method.

Fig. 9(a) and (b) depict the analysis results of influence of window size and minimum support threshold vs. average number of outliers detected in vital dataset. It is observed that the average

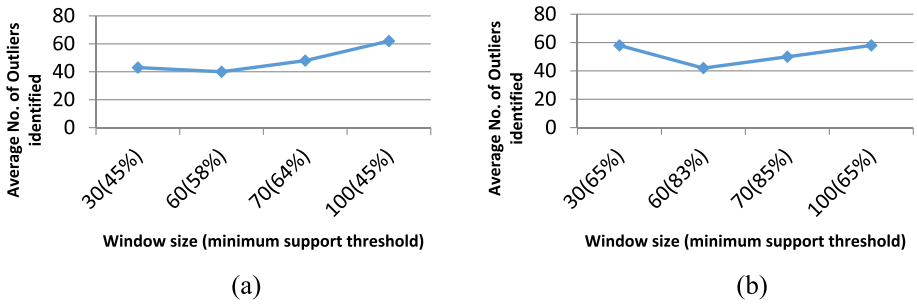

Fig. 10. Average number of outliers vs. window size (minimum support threshold) – Wisconsin breast cancer dataset.

number of identified outliers does not vary with same window size of different minimum support threshold. This is due to fact that the outliers i.e., simulated abnormal data embedded in normal vital dataset are uniformly distributed. In Wisconsin dataset the outliers are randomly distributed and hence the average number of identified outliers slightly differs with same window size of different minimum support threshold as shown in Fig. 10(a) and (b). It can be observed that the final results are stable in both datasets. Therefore, the proposed method is robust to subjective parameters-window size and minimum support threshold.

6. Conclusion

In this paper, a novel technique for detecting outlier transactions in data stream based on minimal infrequent patterns is proposed. A multi pass algorithm MIP-DS is used to mine minimal infrequent patterns. Based on the concept of minimal infrequent patterns, three factors namely TWF, MIPDF and MIFPOF are defined to compute the outlier degree of a transaction. The proposed technique works in two stages. In stage 1, sliding window is adopted to have a bounded dataset and minimal infrequent patterns are mined from the observed stream using MIP-DS algorithm. In stage 2, MIPDF is computed for each minimal infrequent pattern which captures the deviation of the support of from the user defined minimum support threshold. Then the sum of MIPDF of minimal infrequent patterns present in each transaction is computed and weighted using TWF. Finally, MIFPOF is computed for each transaction using TWF and MIPDF. Either top 'k' transactions or transactions having MIFPOF value greater than the outlier threshold are flagged as outliers.

The performance of the proposed MIFPOD method is evaluated using sliding window of varied sizes to handle the data stream and with different minimum support threshold. It is observed from the experimental results of synthetic dataset that the top 'k' outliers detected by MIFPOD method are true outliers whereas in frequent pattern based, distance based and K-nearest neighbor based methods, the detection precision is far less. In the case of real dataset, the top 'k' outliers identified by MIFPOD are more close to FP method but better than distance based and K-nearest neighbor based methods. However, the running time of MIFPOD method is less compared to FP based method. It is due to the fact that the MIFPOD method uses minimal infrequent patterns which are generators of set of all infrequent patterns and hence they form the concise representation of rare patterns that cause the abnormality in the observed data streams opposed to huge number of frequent patterns being used in FP methods. Besides detecting outliers accurately, the proposed method presents the subspace in the form of mined minimal infrequent patterns that cause the transactions to be outliers opposed to distance based and K-nearest neighbor based method which uses full dimension space.

The integration of data mining with outlier detection provides good interpretability of mined results in the form of minimal infrequent patterns and thus facilitates business users to have a good insight of data in a single step. Moreover, the proposed method can extract useful information in sensor data streams for identifying meaning outliers such as abnormality in vital data much faster compared to FP based method.

The proposed method requires transformation of original dataset to binary dataset for further processing which may lead to loss of information. Also, it requires presetting fixed minimum support threshold. In future, it is decided to find a solution for handling continuous data while mining minimal infrequent patterns without the need for binary transformation and to adjust the parameter minimum support threshold automatically to learn concept drift and changes in statistical properties of data stream.

Acknowledgement

This research project is supported by DST-NRDMS, Department of Science and Technology, Government of India, New Delhi. The authors would like to extend their sincere thanks to DST-NRDMS for their support.

References

Adda, M., Wu, L., & Feng, Y. (2007). Rare itemset mining. In *Proceedings of the 6th international conference on machine learning and applications (ICMLA '07)* (pp. 73–80). Washington, DC: IEEE Computer Society.

Aggarwal, C. C., & Philip, S. Y. (2008). Outlier detection with uncertain data. In *SDM5* (pp. 483–493).

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD conference on management of data* (Vol. 22, pp. 207–216). Washington, DC.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, 307–328. AAAI.

Albertini, M. K., & de Mello, R. F. (2007). A self-organizing neural network for detecting novelties. In *Proceedings of the 2007 ACM symposium on applied computing* (pp. 462–466). ACM.

Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European conference on principles of data mining and knowledge discovery (PKDD'02)* (pp. 15–26).

Anguiulli, F., & Fassetti, F. (2007). Detecting distance-based outliers in streams of data. In *Proceedings of the 16th ACM conference on information and knowledge management (CIKM' 07)* (pp. 811–820).

Ashish, G., Akshay, M., & Arnab, B. (2011). Minimally infrequent itemset mining using pattern-growth paradigm and residual trees. In *Proceedings of international conference on management of data (COMAD)* (pp. 57–68).

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. John Wiley and Sons.

Barreto, G. A., & Aguayo, L. (2009). Time series clustering for anomaly detection using competitive neural networks. In *Advances in self-organizing maps* (pp. 28–36). Berlin Heidelberg: Springer.

Böhm, C., Plant, C., Shao, J., & Yang, Q. (2010). Clustering by synchronization. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 583–592). ACM.

Breunig, M. M., Kriegel, H. -P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104). Dallas, USA.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 15.

Chang, J., & Lee, W. (2003). Finding recent frequent itemsets adaptively over online data streams. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 487–492).

Chang, J., & Lee, W. (2004). A sliding window method for finding recently frequent itemsets over online data streams. *Journal of Information Science and Engineering, 20*(4), 753–762.

Chi, Y., Wang, H., Yu, P. S., & Muntz, R. R. (2006). Catch the moment: Maintaining closed frequent itemsets over a data stream sliding window. *Knowledge and Information Systems, 10*(3), 265–294.

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research, 168*(1), 151–168.

Elahi, M., Li, K., Nisar, W., Lv, X., & Wang, H. (2008). Efficient clustering-based outlier detection algorithm for dynamic data stream. *Fifth international conference on fuzzy systems and knowledge discovery, FSKD'08* (Vol. 5, pp. 298–304). IEEE.

Filippone, M., & Sanguinetti, G. (2010). Information theoretic novelty detection. *Pattern Recognition, 43*(3), 805–814.

Fritzke, B. (1995). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems, 7*, 625–632.

GarcíA-RodríGuez, J., Angelopoulou, A., García-Chamizo, J. M., Psarrou, A., Orts Escolano, S., & Morell Gimenez, V. (2012). Autonomous growing neural gas for applications with time constraint: Optimal parameter estimation. *Neural Networks, 32*, 196–208.

Haglin, D. J., & Manning, A. M. (2007). On minimal infrequent itemset mining. In *Proceedings of the international conference DMIN* (pp 141–147). Las Vegas, Nevada, USA.

Hawkins, D. (1980). *Identifications of outliers*. London: Chapman and Hall.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster based local outliers. *Pattern Recognition Letters, 24*(9), 1641–1650.

He, Z., Xu, X., Huang, Z. J., & Deng, S. (2005). FP-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems/ComSIS, 2*(1), 103–118.

Hettich, S., & Bay, S. D. (1999). The UCI KDD repository. Available online at: <http://www.kdd.ics.uci.edu>.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems, 26*(2), 309–336.

Huang, D., Koh, Y. S., & Dobbin, G. (2012). Rare pattern mining on datastream. In *Proceedings of the 14th international conference on data warehousing and knowledge discovery* (pp 303–314).

Ishida, K., & Kitagawa, H. (2008). Detecting current outliers: Continuous outlier detection over time-series data streams. In *Database and expert systems applications* (pp. 255–268). Berlin Heidelberg: Springer.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jiang, S. Y., & An, Q. B. (2008). Clustering-based outlier detection method. *Fifth international conference on fuzzy systems and knowledge discovery, FSKD'08* (Vol. 2, pp. 429–433). IEEE.

Jiang, F., Sui, Y. F., & Cao, C. G. (2010). An information entropy-based approach to outlier detection in rough sets. *Expert Systems with Applications, 37*(9), 6338–6344.

Jiang, M. F., Tseng, S. S., & Su, C. M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters, 22*(6–7), 691–700.

Kao, L. J., & Huang, Y. P. (2012). Association rules based algorithm for identifying outlier transactions in data stream. In *2012 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 3209–3214). IEEE.

Kit, D., Sullivan, B., & Ballard, D. (2011). Novelty detection using growing neural gas for visuo-spatial memory. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1194–1200). IEEE.

Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB conference* (pp. 392–403). New York.

Knorr, E., Ng, R., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *VLDB Journal: Very Large Databases, 8*(3–4), 237–253.

Koh, Y. S., & Rountree, N. (2005). Finding sporadic rules using Apriori-inverse. In *Proceedings of the 9th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD '05)* (pp. 97–106). New York.

Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., & Manolopoulos, Y. (2011). Continuous monitoring of distance-based outliers over data streams. In *2011 IEEE 27th international conference on data engineering (ICDE)* (pp. 135–146). IEEE.

Kriegel, H. P., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 444–452). ACM.

Lee, C.-H., Lin, C.-R., & Chen, M.-S. (2005). Sliding window filtering: An efficient method for incremental mining on a time-variant database. *Information Systems, 30*, 227–244.

Li, H. -F., Lee, S. -Y., Shan, M. -K. (2004). An efficient algorithm for mining frequent itemsets over the entire history of data streams. In *Proceedings of the international workshop on knowledge discovery in data streams*.

Li, Y., Li, D., Wang, S., & Zhai, Y. (2014). Incremental entropy-based clustering on categorical data streams with concept drift. *Knowledge-Based Systems, 59*, 33–47.

Lin, F., Le, W., & Bo, J. (2010). Research on maximal frequent pattern outlier factor for online high-dimensional time-series outlier detection. *Journal of Convergence Information Technology, 5*(10), 66–71.

Liu, J., & Deng, H. (2013). Outlier detection on uncertain data based on local information. *Knowledge-Based Systems, 51*, 60–71.

Liu, B., Hsu, W., & Ma, Y. (1999). Mining association rules with multiple minimum supports. In *Proceedings of 5th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '99)* (pp 337–341). New York.

Liu, B., Xiao, Y., Cao, L., Hao, Z., & Deng, F. (2013). SVDD-based outlier detection on uncertain data. *Knowledge and Information Systems, 34*(3), 597–618.

Manku, G. S., & Motwani, R. (2002). Approximate frequency counts over data streams. In *Proceedings of the VLDB* (pp. 346–357).

Marsland, S., Shapiro, J., & Nehmzow, U. (2002). A self-organising network that grows when required. *Neural Networks, 15*(8), 1041–1058.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing, 99*, 215–249.

Potter, P. A., Perry, A. G., Castaldi, P., Stockert, P., & Hall, A. (2011). *Study guide for basic nursing* (7th ed.). Canada: Elsevier. Vital Signs (pp. 259–308), (Chapter 14).

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large datasets. In *Proceedings of the ACM SIGMOD conference on management of data* (pp. 427–438). Dallas.

Ren, J., Wu, Q., Zhang, J., & Hu, C. (2009). Efficient outlier detection algorithm for heterogeneous data streams. *Sixth international conference on fuzzy systems and knowledge discovery, FSKD'09* (Vol. 5, pp. 259–264). IEEE.

Shaikh, S. A., & Kitagawa, H. (2012). Distance-based outlier detection on uncertain data of Gaussian distribution. In *Web technologies and applications* (pp. 109–121). Berlin Heidelberg: Springer.

Shao, J., Böhm, C., Yang, Q., & Plant, C. (2010). Synchronization based outlier detection. In *Machine learning and knowledge discovery in databases* (pp. 245–260). Berlin Heidelberg: Springer.

Szathmary, L., Napoli, A., & Valtchev, P. (2007). Towards rare itemset mining. *Proceedings of the 19th IEEE international conference on tools with artificial intelligence (ICTAI)* (Vol. 1, pp. 305–312). Los Alamitos: IEEE Computer Society.

Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., & Filzmoser, P. (2013). Locally centred mahalanobis distance: A new distance measure with salient features towards outlier detection. *Analytica Chimica Acta, 787*, 1–9.

Troiano, L., & Scibelli, G. (2014). A time efficient breadth-first level-wise lattice traversal algorithm to discover rare itemsets. *Data Mining and Knowledge Discovery, 28*(3), 773–807.

Tsang, S., Koh, Y. S., & Dobbie, G. (2011). RP-tree: Rare pattern tree mining. In *Proceedings of the 13th international conference on data warehousing and knowledge discovery* (pp. 277–288).

Wei, L., Qian, W., Zhou, A., Jin, W., & Yu, J. X. (2003). HOT: Hypergraph-based outlier test for categorical data. In *Proceedings of the 7th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD '03)* (pp. 399–410).

Wu, Q., & Ma, S. (2011). Detecting outliers in sliding window over categorical data streams. *2011 Eighth international conference on fuzzy systems and knowledge discovery (FSKD)* (Vol. 3, pp. 1663–1667). IEEE.

Wu, S., & Wang, S. (2013). Information-theoretic outlier detection for large-scale categorical data. *IEEE Transactions on Knowledge and Data Engineering, 25*(3), 589–602.

Yamanishi, K., & Takeuchi, J. (2001). Discovering outlier filtering rules from unlabeled data-combining a supervised learner with an unsupervised learner. In *Proceedings of KDD'01* (pp. 389–394).

Yamanishi, K., Takeuchi, J., & Williams, G. (2000). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of KDD'00* (pp. 320–325).

Yan, Q. Y., Xia, S. X., & Feng, K. W. (2012). Probabilistic distance based abnormal pattern detection in uncertain series data. *Knowledge-Based Systems, 36*, 182–190.

Yu, J.-X., Chong, Z., Lu, H., Zhang, Z., & Zhou, A. (2006). A false negative approach to mining frequent itemsets from high speed transactional data streams. *Information Sciences, 176*(14), 1986–2015.

Yu, H., Wang, B., Xiao, G., & Yang, X. (2010). Distance-based outlier detection on uncertain data. *Journal of Computer Research and Development, 47*(3), 474–484.

Zephyr Bio-Harness. <http://www.zephyr-technology.com>.

Zephyr Pressure Monitor <http://www.zephyranywherestore.com/Automatic-Bluetooth-Pressure-Monitor-HPL-108/dp/B009ZUG2Z8>.

Zhang, W., Wu, J., & Yu, J. (2010). An improved method of outlier detection based on frequent pattern. In *Proceedings of WASE international conference on information engineering* (pp. 3–6).

Zhou, X. Y., Sun, Z. H., Zhang, B. L., & Yang, Y. D. (2007). Fast outlier detection algorithm for high dimensional categorical data streams. *Ruan Jian Xue Bao (Journal of Software), 18*(4), 933–942.