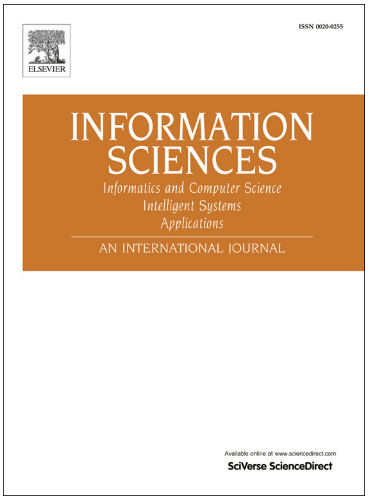
Trend analysis of categorical data streams with a concept change method

Fuyuan Cao, Joshua Zhexue Huang, Jiye Liang

PII: S0020-0255(14)00158-3
DOI: http://dx.doi.org/10.1016/j.ins.2014.02.052
Reference: INS 10569

To appear in: *Information Sciences*



ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES

Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Available online at www.sciencedirect.com
SciVerse ScienceDirect

Please cite this article as: F. Cao, J.Z. Huang, J. Liang, Trend analysis of categorical data streams with a concept change method, *Information Sciences* (2014), doi: http://dx.doi.org/10.1016/j.ins.2014.02.052

Trend analysis of categorical data streams with a concept change method

Fuyuan Cao[a,c], Joshua Zhexue Huang[b,a,*], Jiye Liang[c]

[a]*Shenzhen Key Laboratory of High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[b]*College of Computer Sciences & Software Engineering, Shenzhen University, Shenzhen 518060, China*
[c]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*

Abstract

This paper proposes a new method to trend analysis of categorical data streams. A data stream is partitioned into a sequence of time windows and the records in each window are assumed to carry a number of concepts represented as clusters. A data labeling algorithm is proposed to identify the concepts or clusters of a window from the concepts of the preceding window. The expression of a concept is presented and the distance between two concepts in two consecutive windows is defined to analyze the change of concepts in consecutive windows. Finally, a trend analysis algorithm is proposed to compute the trend of concept change in a data stream over the sequence of consecutive time windows. The methods for measuring the significance of an attribute that causes the concept change and the outlier degrees of objects are presented to reveal the causes of concept change. Experiments on real data sets are presented to demonstrate the benefits of the trend analysis method.

Keywords: Trend analysis, Categorical data stream, Concept change

[*] Corresponding author

Email addresses: `cfy@sxu.edu.cn` (Fuyuan Cao), `zx.huang@szu.edu.cn` (Joshua Zhexue Huang), `ljy@sxu.edu.cn` (Jiye Liang)

1. Introduction

Many real world applications generate continuously arriving data, such as business transactions, network event logs and social networks. This type of data is known as data streams [23]. In data stream mining, most research has been focused on numerical data streams [1, 2, 5, 12, 13, 18]. Recently, the mining of categorical data streams has become a research topic of growing interest [4, 10, 11, 14, 19, 20].

A data stream can be considered as a sequence of data records, each representing an object with a timestamp. Given a time window, we assume that the objects represented by these records within the time window are distributed in different clusters and each cluster represents a concept. As new data records arrive over time, the structure of clusters changes, which results in change of concepts represented in the clusters. In this context, a concept change is called concept drift [23].

Two types of concept drift are illustrated in [17]. One is sudden (abrupt) concept drift and the other is gradual concept drift. Sudden concept drift is described as that the structure of clusters is changed dramatically in short time. Gradual concept drift is considered that the change of a concept occurs gradually over time. For example, in social network analysis, people in a social group or cluster are interested in a particular topic at certain time period. Some people gradually change their interest in the topic and some suddenly change their interests from the current topic to a new topic. The former represents a gradual concept drift and the later is a sudden concept drift.

In [6], we have defined a difference measure to compute the change of concepts between two consecutive windows. With this measure, we are able to analyze the trend of concept change over time through the change of clusters in consecutive windows. However, this measure cannot reveal the relative concept change between two time windows. To solve this problem, we have defined in [8] the new concept emerging degree and the old concept fading degree to measure the relative concept change between two consecutive windows.

In this paper, we propose a new method to trend analysis of categorical data streams by extending the work of [6] and [8]. In this method, we partition a data stream into a sequence of time windows. The data records in each window are assumed to carry a number of concepts represented as clusters. We propose a data labeling algorithm to identify the concepts or clusters in a window from the concepts of the preceding window. We express concepts following the idea of Node Importance [10] and define the distance between two concepts in two consecutive windows using the new concept emerging degree and the old concept fading degree to analyze the change of concepts in consecutive windows. We present the methods for measuring the significance of an attribute that causes the concept change and the outlier degrees of objects to reveal the causes of concept change. Finally, we integrate the above techniques in a trend analysis algorithm to compute the trend of concept change in a data stream over the sequence of consecutive time windows.

A series of experiments were conducted on KDD-CUP'99 data set [22]. The experimental results have shown that the proposed method can discover the trend of concept change in consecutive windows. In comparison with [6], the new method not only revealed the relative concept change between consecutive windows but also found the causes of concept change.

The rest of this paper is organized as follows. Section 2 states the research problem. Section 3 reviews some preliminaries. The trend analysis algorithm and the corresponding techniques are presented in Section 4. Experimental results on real data sets are shown in Section 5. The paper is concluded in Section 6.

2. Problem Statement

A categorical data stream consists of a sequence of records or objects with timestamps, where each record is described by a set of categorical attributes such as Sex, Position, Location and Class. A categorical attribute takes values from a finite set of categories, for instance, Sex=$\{M, F\}$. Formally, a categorical data stream can be formulated as a table of the quintuple $TDT = (U, A, V, f, t)$, where U is a nonempty set of objects called the universe, A is a nonempty set of attributes, t is a sequence of timestamps, $f : U \times A \times t \to V$ is a mapping called an information function such that for any $x \in U$, $a \in A$ and $t' \in t$, $f(x, a, t') \in V_a$, where V_a is a finite and unordered set of values for attribute a. $V = \bigcup_{a \in A} V_a$ is the union of all attribute domains.

Given a particular categorical data stream, we partition the sequence of objects into a set of consecutive time windows with respect to t, using the sliding window technique[3, 9, 16]. Suppose that N is the size of a sliding window, i.e., the number of records in the window, data stream TDT is partitioned into a series of subsets $S^{T_i}(1 \leq i \leq \lceil \frac{|U|}{N} \rceil)$, where T_i represents the ith window and $S^{T_i} \bigcap S^{T_j} = \emptyset$ $(1 \leq i \neq j \leq \lceil \frac{|U|}{N} \rceil)$.

Problem Statement: Given a categorical data stream whose objects are partitioned into a set of consecutive time windows, find the concepts the objects carry in each window; find the change of concepts in two consecutive windows; find the causes of concept change; find the trend of concept change over the sequence of consecutive time windows.

3. Preliminaries

In this section, we briefly review some definitions, such as the new concept emerging degree, the old concept fading degree and the difference measure between two windows that are used to measure concept change. These definitions were first given in [8].

Definition 1. *[21] Let $TDT = (U, A, V, f, t)$ be a categorical data stream, $P \subseteq A$ and $X \subseteq U$. For any $Y \subseteq X$ and $x \in X$, the lower approximation and upper approximation of Y in X are defined as*

$$\underline{P}Y = \{x | [x]_P \subseteq Y\} \tag{1}$$

and

$$\overline{P}Y = \{x | [x]_P \bigcap Y \neq \emptyset\}, \tag{2}$$

where $[x]_P = \{y \in X | (x, y) \in IND(P)\}$. $IND(P)$ is an equivalence relation, which is defined as $IND(P) = \{(x, y) \in X \times X | \forall a \in P, f(x, a) = f(y, a)\}$.

Here, we describe the lower approximation and upper approximation of Y in a set X, not the universe U.

Given a categorical data stream whose objects are partitioned into a set of consecutive windows, we can use Definition 1 to measure the change of concepts between two consecutive windows. For example, in a social media data stream, a time window may contain several topics (concepts). The set of topics change as a new topic emerges in the following window or an old topic disappears. The intuitive example in Figure 1 illustrates three types of concept change. Assume the two rectangles in each subfigure represent two consecutive windows. The circles in each window indicate different concepts. Figure 1(a) shows the concept described by the yellow circle emerged in the following window. Figure 1(b) shows the concept described by the yellow circle disappeared in the following window. In Figure 1(c), two old concepts faded completely and two new concepts emerged in the following window.

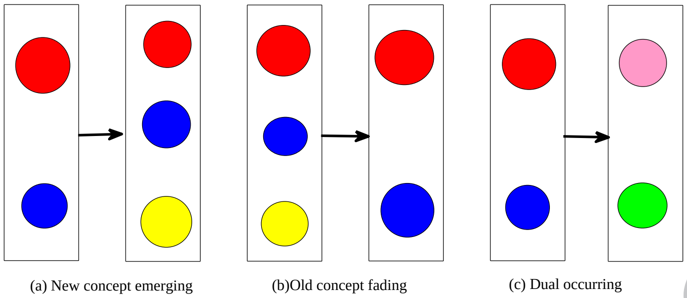

(a) New concept emerging (b)Old concept fading (c) Dual occurring

Figure 1: Three types of concept change

Using the lower approximation and upper approximation in Definition 1, we define the new concept emerging degree and the old concept fading degree in two consecutive windows as follows.

Definition 2. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. The new concept emerging degree and the old concept fading degree from T_i to T_j with respect to A are defined as*

$$NED_A < S^{T_i}, S^{T_j} > \ = \tfrac{1}{|A|} \textstyle\sum_{a \in A} NED_{\{a\}} < S^{T_i}, S^{T_j} > \tag{3}$$

and

$$OFD_A < S^{T_i}, S^{T_j} > \ = \tfrac{1}{|A|} \textstyle\sum_{a \in A} OFD_{\{a\}} < S^{T_i}, S^{T_j} >, \tag{4}$$

where

$$NED_{\{a\}} < S^{T_i}, S^{T_j} >= \frac{|\underline{\{a\}}S^{T_j}|}{|\overline{\{a\}}S^{T_j}|},$$

$$OFD_{\{a\}} < S^{T_i}, S^{T_j} >= \frac{|\underline{\{a\}}S^{T_i}|}{|\overline{\{a\}}S^{T_i}|}.$$

Here, $\underline{\{a\}}S^{T_m}$ and $\overline{\{a\}}S^{T_m}(m = i, j)$ represent the lower approximation and the upper approximation of S^{T_m} in $S^{[T_i,T_j]}$ with respect to attribute a, respectively. The objects in $\underline{\{a\}}S^{T_m}$ can be with certainty classified as members of S^{T_m} on the basis of knowledge in a, while the objects in $\overline{\{a\}}S^{T_m}$ can be only classified as possible members of S^{T_m} on the basis of knowledge in a.

$NED_A < S^{T_i}, S^{T_j} >$ represents the accuracy of approximation [21] of S^{T_j} in $S^{[T_i,T_j]}$, while $OFD_A < S^{T_i}, S^{T_j} >$ is the accuracy of approximation of S^{T_i} in $S^{[T_i,T_j]}$. The higher the two measures, the bigger the relative concept change occurring in the two windows. That is to say, the higher the values of $NED_A < S^{T_i}, S^{T_j} >$ or $OFD_A < S^{T_i}, S^{T_j} >$ are, the bigger the difference between S^{T_i} and S^{T_j}.

If $S^{[T_i,T_j]}/IND(\{a\}) = \{X | X = \{u\}, u \in S^{[T_i,T_j]}\}$, $NED_{\{a\}} < S^{T_i}, S^{T_j} >$ and $OFD_{\{a\}} < S^{T_i}, S^{T_j} >$ achieve their maximum value 1. In other words, $NED_{\{a\}} < S^{T_i}, S^{T_j} >$ and $OFD_{\{a\}} < S^{T_i}, S^{T_j} >$ are precise with respect to a.

If $S^{[T_i,T_j]}/IND(\{a\}) = \{X | X = S^{[T_i,T_j]}\}$, $NED_{\{a\}} < S^{T_i}, S^{T_j} >$ and $OFD_{\{a\}} < S^{T_i}, S^{T_j} >$ achieve their minimum value 0. In other words, $NED_{\{a\}} < S^{T_i}, S^{T_j} >$ and $OFD_{\{a\}} < S^{T_i}, S^{T_j} >$ are vague with respect to a.

Obviously, we have $0 \leq NED_A < S^{T_i}, S^{T_j} > \leq 1$ and $0 \leq OFD_A < S^{T_i}, S^{T_j} > \leq 1$.

Figure 1 shows that old concept fading and new concept emerging can occur simultaneously. We use $NED_A < S^{T_i}, S^{T_j} >$ and $OFD_A < S^{T_i}, S^{T_j} >$ to define the difference measure between two consecutive windows as follows.

Definition 3. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. The difference measure between S^{T_i} and S^{T_j} with respect to A is defined as*

$$DM_A(S^{T_i}, S^{T_j}) = \frac{NED_A < S^{T_i}, S^{T_j} > + OFD_A < S^{T_i}, S^{T_j} >}{2}. \quad (5)$$

We can verify that $DM_A(S^{T_i}, S^{T_j})$ is a distance metric.

4. Trend Analysis Method

In this section, we propose a new algorithm for trend analysis of concept change in categorical data streams. We first present a data labeling algorithm to identify concepts of a given time window from the concepts in the preceding window. Then, we define a method to express concepts and a distance measure of two concepts in two consecutive windows. After that, we integrate all these methods in the trend analysis algorithm. Finally, we present methods to measure the significance of an attribute that affects the concept change and the outlier degree of objects in a time window.

4.1. Data-labeling algorithm

Given the set of objects in the first window of a data stream, we can use a clustering algorithm to divide the objects into clusters and identify concepts. If the difference measure between T_i and T_j is greater than a given threshold, we consider that T_j is a concept-drifting window relative to T_i and use a clustering algorithm to find new concepts in T_j. If there is no significant change in concepts between two consecutive windows, we can use a data labeling method to quickly partition the objects in the current window by referencing the concepts in the preceding window. Inspired by the idea of Node Importance [10], we define the degree of membership of an object in the current window T_j that belongs to a cluster or concept in the preceding window T_i as follows.

Definition 4. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. Suppose that $C^{T_i} = \{c_1^{T_i}, c_2^{T_i}, \cdots, c_{k_{T_i}}^{T_i}\}$ is the clustering results on S^{T_i}, where $c_m^{T_i}$ is the mth cluster, $1 \le m \le k_{T_i}$. For any unlabeled object $x \in S^{T_j}$, the degree of membership of x belonging to $c_m^{T_i}$ with respect to A is defined as*

$$Sim_A(x, c_m^{T_i}) = \sum_{a \in A} \delta_a \times \omega_a, \tag{6}$$

where

$$\delta_a = \frac{|\{y|f(x,a) = f(y,a), y \in c_m^{T_i}\}|}{|c_m^{T_i}|}$$

and

$$\omega_a = 1 + \frac{1}{log_2(k_{T_i})} \times \sum_{m=1}^{k_{T_i}} (q_a \times log_2(q_a)).$$

The value of δ_a reflects the frequency of the component $f(x,a)$ in $c_m^{T_i}$. In other words, the component is important in the cluster when the frequency of the component is high in this cluster. The value of ω_a measures the entropy of component $f(x,a)$ between clusters, where $q_a = \frac{|\{y|f(x,a)=f(y,a), y \in c_m^{T_i}\}|}{|\{z|f(z,a)=f(x,a), z \in C^{T_i}\}|}$. Suppose that there is a component which occurs in all clusters uniformly, the component which contains the maximum uncertainty provides less similarity. In other words, attribute a is of no effect for the degree of membership.

Table 1: An example of categorical data stream

Object	A_1	A_2	A_3	A_4
x_1	A	M	C	t_1
x_2	Y	E	P	t_2
x_3	X	E	P	t_3
x_4	Y	M	P	t_4
x_5	A	M	D	t_5
x_6	A	M	C	t_6
x_7	X	M	P	t_7
x_8	A	M	D	t_8
x_9	Y	M	P	t_9
x_{10}	A	M	C	t_{10}
x_{11}	B	E	G	t_{11}
x_{12}	X	M	P	t_{12}
x_{13}	B	E	D	t_{13}
x_{14}	Y	M	P	t_{14}
x_{15}	B	F	D	t_{15}
x_{16}	Y	M	P	t_{16}
x_{17}	X	M	P	t_{17}
x_{18}	Z	N	T	t_{18}
x_{19}	X	M	P	t_{19}
x_{20}	Y	M	P	t_{20}

We use an example to show that $DM_A()$ can measure not only the difference between two windows, but also the relative concept change between two windows. The data is shown in Table 1.

Let $U = \{x_1, x_2, \cdots, x_{20}\}$, $A = \{A_1, A_2, A_3, A_4\}$, where A_4 is the timestamp. Suppose that the size of the time window is 5. We have $S^{T_1} = \{x_1, x_2, \cdots, x_5\}$, $S^{T_2} = \{x_6, x_7, \cdots, x_{10}\}$, $S^{T_3} = \{x_{11}, x_{12}, \cdots, x_{15}\}$ and $S^{T_4} = \{x_{16}, x_{17}, \cdots, x_{20}\}$. Using Definition 3, we have $DM_A(S^{T_1}, S^{T_2}) = 0.0333$, $DM_A(S^{T_2}, S^{T_3}) = 0.2507$ and $DM_A(S^{T_3}, S^{T_4}) = 0.2381$. We set the threshold of concept drift to 0.2. Since $DM_A(S^{T_1}, S^{T_2}) \leq 0.2$, we have to allocate the most appropriate cluster label to each object of S^{T_2}. We first used the k-modes algorithm [15] to partition S^{T_1}. Assume that x_1, x_2 were chosen as the initial cluster centers in S^{T_1}. We obtained the clustering results $C^{T_1} = \{c_1^{T_1}, c_2^{T_1}\}$, where $c_1^{T_1} = \{x_1, x_5\}$ and $c_2^{T_1} = \{x_2, x_3, x_4\}$. Table 2 shows the degree of membership between each object in S^{T_2} and each cluster in S^{T_1} according to Definition 4.

Table 2: The degrees of membership between objects of S^{T_2} and clusters of S^{T_1}

	x_6	x_7	x_8	x_9	x_{10}
$c_1^{T_1}$	1.5817	0.0817	1.5817	0.0817	1.5817
$c_2^{T_1}$	0.0272	1.3606	0.0272	1.6939	0.0272

From Table 2, we can obtain that $c_1^{T_2} = \{x_6, x_8, x_{10}\}$ and $c_2^{T_2} = \{x_7, x_9\}$.

The data labeling algorithm is described in *Algorithm* 1. The time complexity for computing the degree of membership between an object and a cluster is $O(|S^{T_i}||A|)$. The total computational cost of the algorithm is $O(|S^{T_i}||A||S^{T_j}|k_{T_i})$. Therefore, this algorithm is linear to the number of the objects in S^{T_j}, i.e., the size of the time window.

Algorithm 1 The data labeling algorithm

```
Input:
- $C^{T_i}$ : the clustering results in $T_i$;
- $S^{T_j}$ : the objects in $T_j$;
Output: a partition of $S^{T_j}$;
Method:
Generate a partition $C^{T_i} = \{c_1^{T_i}, c_2^{T_i}, \cdots, c_{k_{T_i}}^{T_i}\}$ of $S^{T_i}$ with respect to $A$
  by calling the corresponding categorical clustering algorithm;
for $j' = 1$ to $|S^{T_j}|$ do
    for $i' = 1$ to $k_{T_i}$ do
        Calculate $Sim_A(x_{j'}, c_{i'}^{T_i})$ according to Definition 4, where $x_{j'}$ is the
          $j'$th object in $S^{T_j}$.
    end for
    Give label $L$ to $x_{j'}$, where $L = \arg\max_{i'=1,\cdots,k_{T_i}} \{Sim_A(x_{j'}, c_{i'}^{T_i})\}$;
end for
Return $C^{T_j} = \{c_1^{T_j}, c_2^{T_j}, \cdots, c_{k_{T_i}}^{T_j}\}$;
```

4.2. Expression of Concepts

The cluster expressions contribute to the understanding of concepts. The "modes" [15] are a traditional expression of clusters for categorical data.

However, “modes” are mainly focused on the intra-cluster similarity and do not take the inter-cluster similarity into account. To solve this problem, we define a new cluster expression that considers both intra- and inter-cluster similarities.

Definition 5. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^T \subseteq U$. Suppose that $C^T = \{c_1^T, c_2^T, \cdots, c_{k_T}^T\}$ is the clustering results on S^T. The expression of $c_i^T \in C^T$ is defined as*

$$R(c_i^T) = \{q_j | q_j = \arg\max_{q_{j'} \in V_{a_j}} \delta'_{a_j} \times \omega'_{a_j}, j = 1, 2, \cdots, |A|\}, \tag{7}$$

where

$$\delta'_{a_j} = \frac{|\{x | f(x, a_j) = q_{j'}, x \in c_i^T\}|}{|c_i^T|}$$

and

$$\omega'_{a_j} = 1 + \frac{1}{log_2(k_T)} \times \psi.$$

Here

$$\psi = \sum_{i=1}^{k_T} \left(\frac{|\{x | f(x) = q_j^{'}, x \in c_i^T\}|}{|\{z | f(z) = q_j^{'}, z \in C^T\}|} \times log_2 \frac{|\{x | f(x) = q_j^{'}, x \in c_i^T\}|}{|\{z | f(z) = q_j^{'}, z \in C^T\}|} \right).$$

Similar to Definition 4, the value of δ'_{a_j} reflects the frequency of $q_j^{'}$ in c_i^T. The value of ω'_{a_j} measures the entropy of component $q_j^{'}$ between clusters.

Continuing from Example 1, we have $c_1^{T_1} = \{x_1, x_5\}$, $c_2^{T_1} = \{x_2, x_3, x_4\}$, $c_1^{T_2} = \{x_6, x_8, x_{10}\}$ and $c_2^{T_2} = \{x_7, x_9\}$. With Definition 5, we can obtain the expression of each cluster as shown in Table 3.

Table 3: The expressions of clusters in 2 consecutive windows

Clusters	Cluster expression
$c_1^{T_1} = \{x_1, x_5\}$	$R(c_1^1) = \{A, M, C\}$
$c_2^{T_1} = \{x_2, x_3, x_4\}$	$R(c_2^1) = \{Y, E, P\}$
$c_1^{T_2} = \{x_6, x_8, x_{10}\}$	$R(c_1^2) = \{A, M, C\}$
$c_2^{T_2} = \{x_7, x_9\}$	$R(c_2^2) = \{X, M, P\}$

4.3. Distance between two concepts in consecutive windows

With the difference measure in Definition 3, we define a new distance between two concepts (clusters) in consecutive windows as follows.

Definition 6. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. Suppose that $C^{T_i} = \{c_1^{T_i}, c_2^{T_i}, \cdots, c_{k_{T_i}}^{T_i}\}$ and $C^{T_j} = \{c_1^{T_j}, c_2^{T_j}, \cdots, c_{k_{T_j}}^{T_j}\}$ are the clustering results on S^{T_i} and S^{T_j}, respectively. The distance between $c_{i'}^{T_i}$ and $c_{j'}^{T_j}$ with respect to A is defined as*

$$d_A(c_{i'}^{T_i}, c_{j'}^{T_j}) = \frac{NED_A < c_{i'}^{T_i}, c_{j'}^{T_j} > +OFD_A < c_{i'}^{T_i}, c_{j'}^{T_j} >}{2}, \tag{8}$$

where $1 \le i' \le k_{T_i}$, $1 \le j' \le k_{T_j}$.

Continuing from Example 1, we have $DM_A(S^{T_2}, S^{T_3}) > 0.2$ and $DM_A(S^{T_3}, S^{T_4}) > 0.2$. We consider that T_3 and T_4 are two concept-drifting windows. Suppose that the clustering results of S^{T_3} and S^{T_4} are $C^{T_3} = \{c_1^{T_3}, c_2^{T_3}\}$ and $C^{T_4} = \{c_1^{T_4}, c_2^{T_4}\}$, where $c_1^{T_3} = \{x_{11}, x_{13}, x_{15}\}$, $c_2^{T_3} = \{x_{12}, x_{14}\}$, $c_1^{T_4} = \{x_{16}, x_{17}, x_{19}, x_{20}\}$ and $c_2^{T_4} = \{x_{18}\}$. With Definition 6, we can compute the distances of clusters in S^{T_i} and $S^{T_{i+1}}$ $(1 \le i \le 3)$ as shown in Table 4.

Table 4: The distances between clusters in two consecutive windows of 4 windows

	$c_1^{T_1} = \{x_1, x_5\}$	$c_2^{T_1} = \{x_2, x_3, x_4\}$
$c_1^{T_2} = \{x_6, x_8, x_{10}\}$	0	0.7222
$c_2^{T_2} = \{x_7, x_9\}$	0.6667	0.0667
	$c_1^{T_2} = \{x_6, x_8, x_{10}\}$	$c_2^{T_2} = \{x_7, x_9\}$
$c_1^{T_3} = \{x_{11}, x_{13}, x_{15}\}$	0.7750	1
$c_2^{T_3} = \{x_{12}, x_{14}\}$	0.6667	0
	$c_1^{T_3} = \{x_{11}, x_{13}, x_{15}\}$	$c_2^{T_3} = \{x_{12}, x_{14}\}$
$c_1^{T_4} = \{x_{16}, x_{17}, x_{19}, x_{20}\}$	1	0
$c_2^{T_4} = \{x_{18}\}$	1	1

4.4. Significance of attributes and outlier degree of objects

To find the causes of concept change in consecutive windows, we measure the significance of an attribute the change of whose values affects the change of clusters in the following window. If the value distributions of an attribute remain the same in the two consecutive windows, then this attribute has little effect on the concept change. The significance of an attribute is measured as follows.

Definition 7. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. For any $a \in A$, the significance of a between S^{T_i} and S^{T_j} is defined as*

$$Sig_{\{a\}}(S^{T_i}, S^{T_j}) \ = \tfrac{DM_{\{a\}}(S^{T_i},S^{T_j})}{\sum_{c\in A} DM_{\{c\}}(S^{T_i},S^{T_j})}. \tag{9}$$

Continuing from Example 1, we can use Definition 7 to compute the significance of each attribute in two consecutive windows as shown in Table 5.

Table 5: The significance of each attribute in two consecutive windows of 4 windows

windows	A_1	A_2	A_3
$T_1 \longrightarrow T_2$	0	1	0
$T_2 \longrightarrow T_3$	0.5698	0.1994	0.2308
$T_3 \longrightarrow T_4$	0.3333	0.3333	0.3333

From Table 5, we can see that A_1 and A_3 has no effect on concept change from T_1 to T_2. A_1, A_2 and A_3 have the same contributions for concept change from T_3 to T_4.

Similarly, different objects provide different contributions for the concept change. If the attribute values of an object rarely occur in two consecutive windows, the object provides the maximal contribution to concept change and can be considered as an outlier [24]. We measure the degree of an object as an outlier as follows.

Definition 8. *Let $TDT = (U, A, V, f, t)$ be a categorical data stream and $S^{T_i}, S^{T_j} \subseteq U$, where $S^{T_i} \bigcap S^{T_j} = \emptyset$ and $S^{[T_i,T_j]} = S^{T_i} \bigcup S^{T_j}$. For any $x \in S^{[T_i,T_j]}$, the outlier degree of the object x with respect to A is defined as*

$$OD_A(x) \ = \tfrac{1}{|A|} \sum_{a\in A}(1 + \tfrac{w(a)}{|S^{[T_i,T_j]}|} \times log_2 \tfrac{w(a)}{|S^{[T_i,T_j]}|}), \tag{10}$$

where $w(a) = |\{z|f(x,a) = f(z,a), z \in S^{[T_i,T_j]}\}|$.

Table 6: The outlier degree of each object in two consecutive windows of 4 windows

$T_1 \longrightarrow T_2$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
	0.5950	0.8157	0.8434	0.5950	0.6226	0.5950	0.6226	0.6226	0.5950	0.5950
$T_2 \longrightarrow T_3$	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
	0.7350	0.7025	0.7074	0.7025	0.7350	0.9322	0.7025	0.8845	0.7025	0.9046
$T_3 \longrightarrow T_4$	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
	0.9322	0.6410	0.9122	0.6410	0.9322	0.6410	0.6410	1.0000	0.6410	0.6410

Continuing from Example 1, the outlier degree of each object in $T_1 \longrightarrow T_2$, $T_2 \longrightarrow T_3$ and $T_3 \longrightarrow T_4$ is shown in Table 6.

From Table 6, we can see that objects x_3, x_{11} and x_{18} have the maximum outlier degree in $T_1 \longrightarrow T_2$, $T_2 \longrightarrow T_3$ and $T_3 \longrightarrow T_4$, respectively.

4.5. Trend analysis algorithm

Integrating the techniques discussed in the previous sections, we define the trend analysis algorithm in *Algorithm* 2. The total computational cost of this algorithm is $O(|S^{T_i}||A|k_{T_i} + |S^{T_j}||A|k_{T_j} + k_{T_i}k_{T_j}|S^{T_i}\bigcup S^{T_j}||A|) = O(k_{T_i}k_{T_j}|S^{T_i}\bigcup S^{T_j}||A|)$.

We use the trend analysis algorithm to analyze the trend of concept change in the data stream of Table 1. We set the threshold γ to 0.2. The trend of concept change in 4 consecutive time windows is shown in Figure 2. The horizontal axis is consecutive time windows. The blue and red circles in each column indicate the clusters in the time window. The size of the circle represents the number of objects. The content in each circle is the expression of concept in each cluster. Similar concepts are linked with the green lines. From this figure, we can understand how concepts change in consecutive windows.

In comparison with the result in [6], Figure 3 shows the relative concept change between windows. We computed the new concept emerging degree and the old concept fading degree in consecutive time windows as shown in Figure 3. We can see that concept change was caused by emerging new concepts or fading old concepts or both. From T_2 to T_3, the new concept emerging degree was greater than the old concept fading degree. This indicates that more new concepts emerged than old concepts faded. However, from T_3 to T_4, more old concepts faded than new concepts emerged. This

Algorithm 2 The trend analysis algorithm

```
Input:
- $C^{T_i}$ : the clustering results in $T_i$;
- $C^{T_j}$ : the clustering results in $T_j$;
- $\gamma$ : the specified threshold;
Output: the trend of concept change from $T_i$ to $T_j$;
Method:
Obtain clustering results $C^{T_i} = \{c_1^{T_i}, c_2^{T_i}, \cdots, c_{k_{T_i}}^{T_i}\}$ and $C^{T_j} = \{c_1^{T_j}, c_2^{T_j}, \cdots, c_{k_{T_j}}^{T_j}\}$ with respect to $A$;
for $i' = 1$ to $k_{T_i}$ do
    Generate $R(c_{i'}^{T_i})$ according to Definition 5;
end for
for $j' = 1$ to $k_{T_j}$ do
    Generate $R(c_{j'}^{T_j})$ according to Definition 5;
end for
for $i' = 1$ to $k_{T_i}$ do
    for $j' = 1$ to $k_{T_j}$ do
        if $d_A(c_{i'}^{T_i}, c_{j'}^{T_j}) \le \gamma$ then
            Connect $c_{i'}^{T_i}, c_{j'}^{T_j}$ with line;
        end if
    end for
end for
```

phenomenon was caused by the fact that a new cluster of x_{11}, x_{13} and x_{15} emerged in T_3 from T_2 and an old cluster of x_{11}, x_{13} and x_{15} in T_3 faded in T_4.

5. Experimental Results

A series of experiments was conducted on real data for evaluation of the proposed trend analysis algorithm. In this section, we present the results of trend analysis on a real data stream for network intrusion detection and investigate the causes of concept change through significance of attributes and discuss the property of parameter γ.

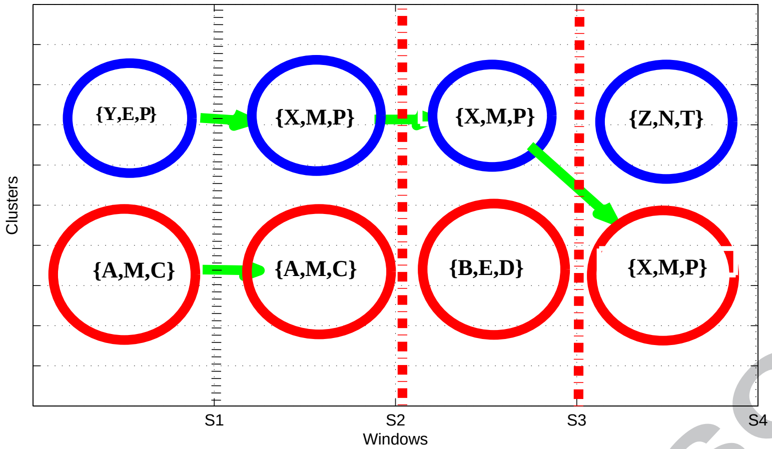


Figure 2: The trend of concept change in 4 consecutive time windows

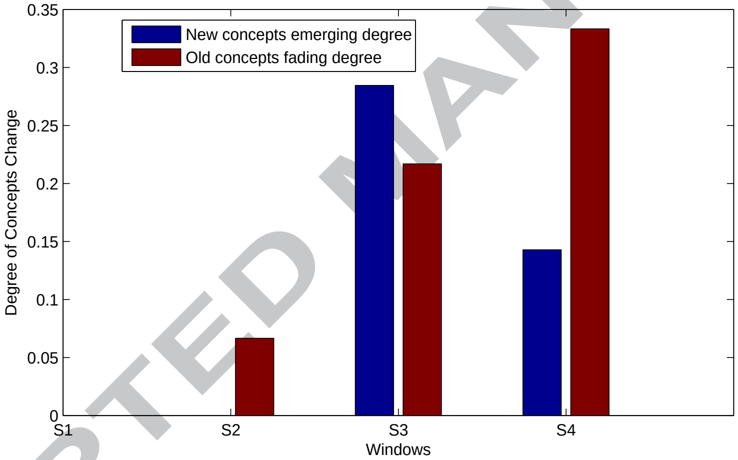


Figure 3: The change of two measures in 4 consecutive windows

5.1. Network stream data

KDD-CUP'99 was used as a test data for The Third International Knowledge Discovery and Data Mining Tools Competition. The data set contained 494,021 records, each having a timestamp. The records were classified into 23 classes. One class indicated the normal connection and other 22 classes were network attack types. Each record was described by 41 attributes, in which 34 attributes were continuous and 7 were categorical. We used uniform quantization to convert these continuous attributes into discrete values, each attribute with 5 categories. We also aggregated 22 attack classes into one general attack class.

5.2. Trends analysis

The first 15000 records in the network data set were selected as a sample data to show trend analysis. We choose 3000 records as the size of the time window and divided the sample data into 5 consecutive time windows. We first used the k-modes algorithm [15] to cluster the records in the first window into two clusters, each representing a concept. Before executing the k-modes algorithm, we used the method in [7] to obtain its initial cluster centers. Then, we used DM_A distance measure Eq.(5) to compute the distance between the first window and the second window. If the distance was smaller than the given threshold 0.01, the data labeling algorithm was used to obtain the concepts for the second window. Otherwise, the k-modes algorithm was used to generate clusters for the second window. This process was repeatedly used to generate concepts in the following windows.

To investigate the relationships of concepts in two consecutive windows, we used d_A distance measure Eq.(8) to compute the similarity between two concepts in the consecutive windows. If the similarity between two concepts was greater than the threshold γ, the two concepts in the consecutive windows were related, i.e., being the same. Figure 4 shows relations of concepts in consecutive windows and the trend of concept change in 5 consecutive time windows. The result was produced with $\gamma = 0.01$. The red circles represent attack clusters and the blue circles are normal connection clusters. The vertical dot lines indicate the boundaries between consecutive time windows. The number in the circle is the number of the records in the cluster. We can see that attacks suddenly emerged in window 3, continued to window 4 and suddenly dropped in window 5. Such trend can help us easily understand the behavior of network attacks over time.

Table 7 shows the expressions of concepts (clusters) of 41 attributes in 5 consecutive windows. Each window has two concepts.

In addition, we compared the proposed method with the work in [6]. In the method of [6], we set $\gamma = 0.1$. Figure 5 shows the trend of concept change in 5 consecutive time windows.

In Figure 5, we find that the clusters between T_3 and T_4 were not connected by lines. In fact, the clusters between T_3 and T_4 should be connected

Table 7: The cluster expressions in 5 consecutive time windows

Attr	$R(c_1^1)$	$R(c_2^1)$	$R(c_1^2)$	$R(c_2^2)$	$R(c_1^3)$	$R(c_2^3)$	$R(c_1^4)$	$R(c_2^4)$	$R(c_1^5)$	$R(c_2^5)$
1	1	1	1	1	1	1	1	1	1	1
2	2	3	2	2	2	2	2	2	1	1
3	20	9	20	20	20	20	20	20	1	1
4	10	10	10	10	10	10	10	10	2	2
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	1
11	0	0	0	0	0	0	0	0	0	0
12	1	0	1	1	1	1	1	1	0	0
13	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1	1
18	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	1	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1	1
29	5	5	5	5	5	5	5	5	2	2
30	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1
32	5	1	1	5	1	5	1	1	1	1
33	5	2	5	5	5	5	5	5	1	1
34	5	5	5	5	5	5	5	5	1	1
35	1	1	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1	1

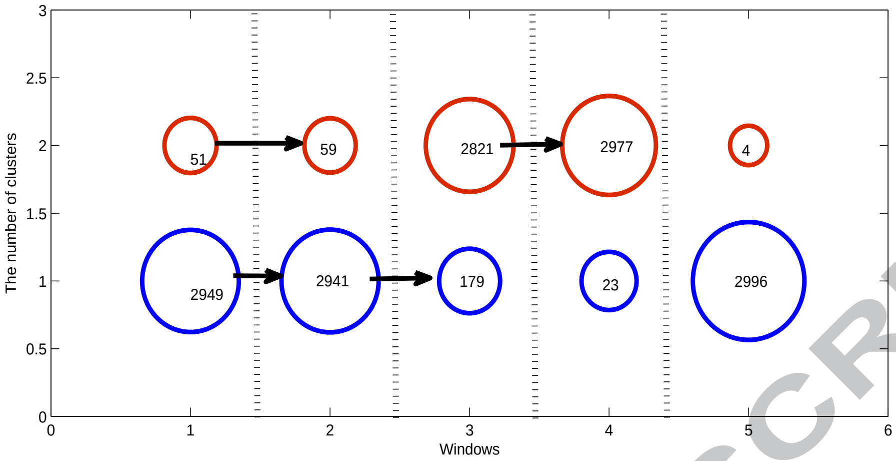


Figure 4: The trend of concept change on the sample set by the proposed method

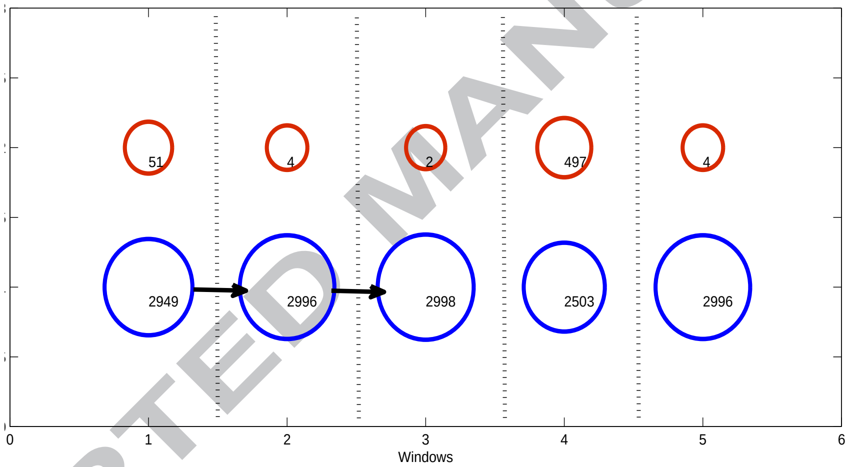


Figure 5: The trend of concept change on the sample set by the method in [6]

because there are many objects labeled by attack in these two time windows. Comparing Figure 4 with Figure 5, we find the results of the proposed method were much closer to the distributions of the sample data.

5.3. Significance of attributes

To investigate the causes of concept change between two windows, we analyzed the significance of attributes for the changed concepts. The result is shown in Table 8. The first column is attribute and the other four columns are the significance measure of each attribute on the transition of two consecutive windows. Value 0 implies that the concept change in consecutive windows was not caused by that attribute. The values greater than 0 indicate that the attributes contributed to the concept change in the consecutive windows.

From Table 8, we can see that a few attributes contributed significantly to the change of concepts in consecutive windows, such as attributes 4, 19, 26, 29, 30, 35, 37 and 38. These attributes were the main causes of the concept change in 5 consecutive windows. Some attributes such as 23 and 24 show significant impact on the concept change in consecutive windows $T_2 \rightarrow T_3$, $T_3 \rightarrow T_4$ and $T_4 \rightarrow T_5$.

Further investigating the causes of concept change, we looked into the value distributions of an attribute in consecutive windows. Table 9 shows the example of attribute 4 in 5 time windows. We can see that the bigger the difference of value distributions in two consecutive windows, the more significant the causes of concept change by the attribute. From this observation, we can monitor the concept change of a data stream by looking into the value change of significant attributes in the data stream.

Computing the new concept emerging degree and the old concept fading degree defined in Definition 2, we investigated the relative concept change in 5 consecutive time windows. The result is shown in Figure 6. We can see a dramatic drop of the fading degree and an obvious rise of the emerging degree from T_4 and T_5. This is an indication that the change was mainly caused by emerging new concepts and the fact was that there were 2488 attack records in T_4 which disappeared in T_5 whereas 3000 normal connection records emerged in T_5.

5.4. Impact of γ

Using the trend analysis algorithm Algorithm 2, we need to specify a threshold γ that determines whether two concepts in consecutive windows are the same concept or one concept in one window has drifted to another concept in the following window. We used the entire data set to investigate γ. The time window size was set as 3000 records and 164 consecutive windows were obtained. We ran the trend analysis algorithm with different values of γ and counted the number of pairs of consecutive windows which had

Table 8: The significance of attributes in 5 consecutive time windows

Attr	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_3$	$T_3 \rightarrow T_4$	$T_4 \rightarrow T_5$
1	0.0345	0.0004	0	0
2	0	0	0	0.2445
3	0	0.0032	0.1195	0.2607
4	0.0345	0.0040	0.0298	0.0008
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0.0690	0	0.0448	0.0003
11	0	0	0.0149	0.0001
12	0	0	0	0
13	0	0	0	0
14	0	0.0016	0	0.0003
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0.0690	0.0008	0	0
19	0.0690	0.0008	0.0149	0.0001
20	0	0	0	0
21	0	0	0	0
22	0	0	0	0.0009
23	0	0.4834	0.1642	0.2319
24	0	0.4834	0.1642	0.2319
25	0.0690	0	0.0298	0.0003
26	0.0690	0.0004	0.0448	0.0001
27	0	0.0016	0	0.0004
28	0	0.0016	0	0.0004
29	0.0345	0.0028	0.0298	0.0025
30	0.0690	0.0028	0.0149	0.0025
31	0	0	0	0
32	0	0	0	0
33	0	0	0	0
34	0	0	0	0.0199
35	0.1034	0.0024	0.0746	0.0012
36	0.2069	0.0032	0.1194	0
37	0.0690	0.0012	0.0149	0.0004
38	0.0690	0.0008	0.0597	0.0002
39	0	0.0016	0.0597	0
40	0	0.0016	0	0.0004
41	0.0345	0.0020	0	0.0004

Table 9: The value distributions of attribute 4

T_1	T_2	T_3	T_4	T_5
{2,7,10}	{2,7,8,10}	{6,10}	{10}	{2,6,7,10}

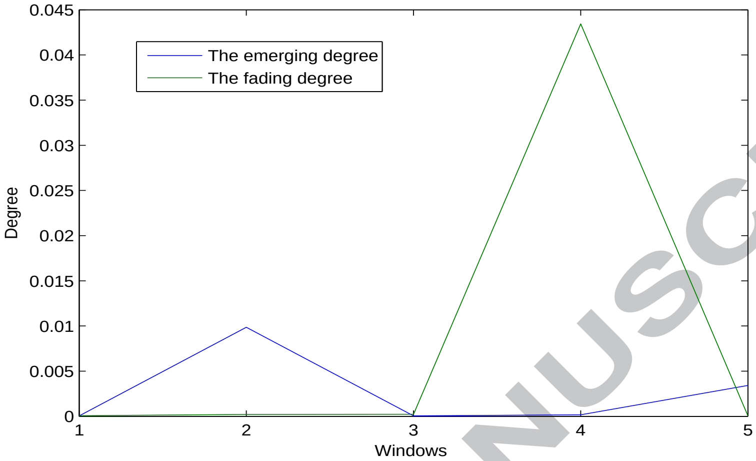


Figure 6: The changes of the new concept emerging degree and the old concept fading degree in 5 consecutive time windows

concept change measured by the distance of two consecutive windows which was greater than γ.

Figure 7 shows the relationships between the number of concept drifts and the value of γ. We can see that the number of drifting-concepts decreases as γ increases. When γ is greater than 0.07, the number of drifting-concepts drops to zero, which means no concept change was identified. Therefore, γ cannot be greater than 0.07 in this data set. To better reveal concept change patterns, we set $\gamma = 0.01$ as default.

6. Conclusions

In this paper, we have presented a new method for trend analysis of categorical data streams. In this method, a data labeling method has been proposed by considering both the intra-cluster similarity and the inter-cluster similarity. We have defined a new distance between concepts in two consecutive windows that is used to measure the concept change. The significance

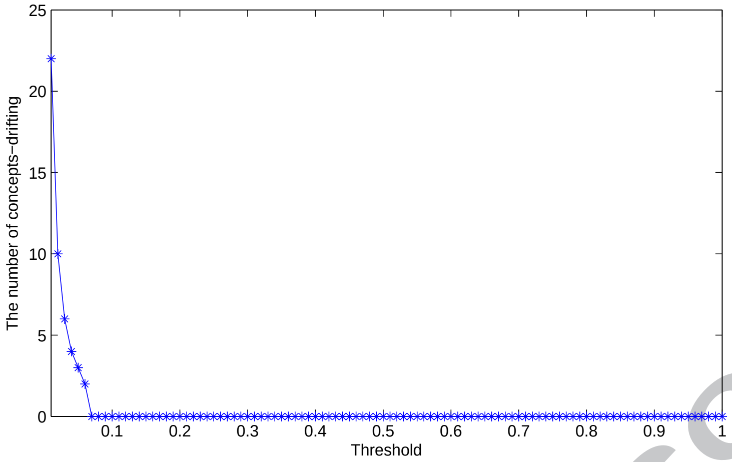


Figure 7: The relationships between the number of concepts-drifting and γ

measure of attributes has also been defined to reveal the causes of concept change. We have used a real data stream to demonstrate the usefulness of the new algorithm in trend analysis.

The trend analysis algorithm proposed in this paper is applied to categorical data streams. Our future work is to study the trend of concept change in the case of continuous data by using the neighborhood rough set because continuous data streams are widely available in real applications.

Acknowledgements

The authors are very grateful to the anonymous reviewers and editor. Their many helpful and constructive comments and suggestions helped us significantly improve this work. This work was supported by Shenzhen Internet Industry Development Foundation (No. JC201005270342A), China Postdoctoral Science Foundation (Nos. 2012M510046, 2013M542219), the National Natural Science Foundation of China (Nos. 71031006, 61305073), the National Key Basic Research and Development Program of China (973) (No. 2013CB329404) and the Technology Research Development Projects of Shanxi (No. 20110002).

References

[1] C. C. Aggarwal, A segment-based framework for modeling and mining data streams, Knowledge and Information Systems 30(1)(2012) 1-29.

[2] C. C. Aggarwal, J. Han, J. Wang, P. S. Yu, A framework for projected clustering of high dimensional data streams, In: Proceedings of Very Large Databases Conference (VLDB) (2004).

[3] C. C. Aggarwal, J. Han, J. Wang, P. S. Yu, A framework for clustering evolving data streams, In: Proceedings of Very Large Databases Conference (VLDB) (2003).

[4] C. C. Aggarwal, P. S. Yu, On clustering massive text and categorical data streams, Knowledge and Information Systems 24(2)(2010) 171-196.

[5] F. Cao, M. Ester, Q. Qian, A. Zhou, Density-based clustering over an evolving data streams with noise, In: Proceedings of SIAM Conference on Data Mining (2006).

[6] F. Y. Cao, J. Y. Liang, L. Bai, X. W. Zhao, C. Y. Dang, A framework for clustering categorical time-evolving data, IEEE Transactions on Fuzzy Systems 18(5)(2010) 872-885.

[7] F. Y. Cao, J. Y. Liang, L. Bai, A new initialization method for categorical data clustering, Expert Systems with Applications 36(7)(2009) 10223-10228.

[8] F. Y. Cao, J. Z. Huang, A concept-drifting detection algorithm for categorical evolving data, In: J. Pei et al.(Eds.): PAKDD 2013, Part II LNAI 7819(2013) 492-503.

[9] D. Chakrabarti, R. Kumar, A. Tomkins, Evolutionary clustering, In: Proceedings of ACM SIGKDD Knowledge Discovery and Data Mining (2006) 554-560.

[10] H. L. Chen, M. S. Chen, S. C. Lin, Catching the trend: A framework for clustering concept-drifting categorical data, IEEE Transactions on Knowledge and Data Engineering 21(5)(2009) 652-665.

[11] K. K. Chen, L. Liu, HE-Tree: a framework for detecting changes in clustering structure for categorical data streams, The VLDB Journal 18(6)(2009) 1241-1260.

[12] Y. Chi, X. D. Song, D. Y. Zhou, K. Hino, B. L. Tseng, Evolutionary spectral clustering by incorporating temporal smoothness, Proceedings of ACM SIGKDD Knowledge Discovery and Data Mining (2007) 153-162.

[13] B. R. Dai, J. W. Huang, M. Y. Yeh, M.S. Chen, Adaptive clustering for multiple evolving steams, IEEE Transactions on Knowledge and Data Engineering 18(9)(2006) 1166-1180.

[14] Z. Y. He, X. F. Xu, S. C. Deng, J. Z. Huang, Clustering categorical data streams, Journal of Computational Methods in Science and Engineering 11(4)(2011) 185-192.

[15] Z. X. Huang, Extensions to the *k*-Means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2(3)(1998) 283-304.

[16] M. M. Gaber, P. S. Yu, Detection and classification of changes in evolving data streams, International Journal of Information Technology and Decision Making 5(4)(2006) 659-670.

[17] S. Guha, A. Meyerson, N. Mishra, R. Motwani, L. OCallaghan, Clustering data streams: theory and practice, IEEE Transactions on Knowledge and Data Engineering 15(3)(2003) 515-528.

[18] N. Mozafari, S. Hashemi, A. Hamzeh, A precise statistical approach for concept change detection in unlabeled data streams, Computers & Mathematics with Applications 62(4)(2011) 1655-1669.

[19] O. Nasraoui, M. Soliman, E. Saka, A. Badia, R. Germain, A web usage mining framework for mining evolving user profiles in dynamic web sites, IEEE Transactions on Knowledge and Data Engineering 20(2)(2008) 202-215.

[20] K. L. Ong, W. Y. Li, W. K. Ng, E. P. Lim, SCLOPE: An algorithm for clustering data streams of categorical attributes, Lecture Notes in Computer Science 3181(2004) 209-218.

[21] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11(5)(1982) 341-356.

[22] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/, 2013.

[23] G. Widmer, M. Kubat, Learning in the presence of concept drift and hidden context, Machine Learning 23(1)(1996) 69-101.

[24] X. W. Zhao, J. Y. Liang, F. Y. Cao, A simple and effective outlier detection algorithm for categorical data. International Journal of Machine Learning and Cybernetics, DOI: 10.1007/s13042-013-0202-4 (2013).