

# Using Fuzzy C-means to Discover Concept-drift Patterns for Membership Functions

Tzung-Pei Hong\* , Chun-Hao Chen , Yan-Kang Li , Min-Thai Wu 

**Abstract.** People often change their minds at different times and at different places. It is important and valuable to indicate concept-drift patterns in unexpected ways for shopping behaviours for commercial applications. Research about concept drift has been growing in recent years. Many algorithms dealt with concept-drift information and detected new market trends. This paper proposes an approach based on fuzzy c-means (FCM) to mine the concept drift of fuzzy membership functions. The proposed algorithm is subdivided into two stages. In the first stage, individual fuzzy membership functions are generated from different training databases by the proposed FCM-based approach. Then, the proposed algorithm will mine the concept-drift patterns from the sets of fuzzy membership functions in the second stage. Experiments on simulated datasets were also conducted to show the effectiveness of the approach.

**AMS Subject Classification 2020:** 03E72; 68T30

**Keywords and Phrases:** Concept drift, Data mining, Fuzzy c-means, Membership function.

## 1 Introduction

Data processing and storage are now more readily available than ever because of the booming development of information technologies. If policy-makers can obtain information and knowledge from databases effectively and quickly, they can make better decisions. However, along with the growing number of database types, getting useful and valuable information from large databases for decision-making is difficult and important [4, 27].

Fuzzy data mining is attracting much research interest these years. In fuzzy data mining, membership functions are given and used to extract fuzzy association rules represented by linguistic terms from quantitative data [3, 8, 9, 11, 26]. Therefore, fuzzy membership functions play a crucial role in affecting the quality of the mining results. In literature, in addition to using static or manually defined membership functions, meta-heuristic methods are utilized to find appropriate membership functions [7, 24]. For example, Hong et al. proposed a genetic-fuzzy mining approach to extract fuzzy association rules with the derived membership functions from given quantitative transactions [7]. Yang et al. proposed a method to generate fuzzy membership functions using unsupervised learning of a self-organizing feature map [24].

This paper discusses the concept-drift issue of membership functions. We present a detection algorithm for concept drift of membership functions. Firstly, each item in the database was divided into several linguistic

\*Corresponding author: Tzung-Pei Hong, Email: [tphong@nuk.edu.tw](mailto:tphong@nuk.edu.tw), ORCID: 0000-0001-7305-6492

Received: 14 May 2022; Revised: 12 June 2022; Accepted: 15 June 2022; Published Online: 7 November 2022.

**How to cite:** T. P. Hong, C. H. Chen, Y. K. Li and M. T. Wu, Using Fuzzy C-means to Discover Concept-drift Patterns for Membership Functions, *Trans. Fuzzy Sets Syst.*, 1(2) (2022), 21-31.

terms, such as low, medium, and high. The proposed FCM-based approach is then used to generate a set of relevant membership functions for each item in two given transactional databases that could be collected at different times or places. Finally, a comparison algorithm is used to obtain the concept-drift patterns for these membership functions in these two databases by a predefined threshold.

The rest of this paper is organized as follows. Section 2 reviews some related researches. Section 3 states three types of concept drift of membership functions. Section 4 describes the proposed algorithm. Section 5 shows the experimental results and analysis. At last, the conclusion and future work are given in Section 6.

## 2 Related Works

In this section, we review some related research in regards to concept drift, fuzzy data mining, FCM, and membership functions.

### 2.1 Concept Drift

The research of concept drift has become popular in recent years [4, 22, 27]. Tsymbal proposed the concept of drift as finding patterns which change over time in unexpected ways [22]. For example, assume at time  $t$  there is an association rule of "if buying milk, then buying bread" and at time  $t + k$  there is another rule "if buying milk, then buying apple". For the two rules, the latter changes from the former in the consequence part along with time. The change is a type of concept-drift pattern.

The traditional methods of data mining have been used in various research areas based on the concept-drift patterns [5, 13, 20, 23]. For instance, in intrusion detection systems, Mukkavilli et al. designed an approach for detecting network attacks [13]. Hayat et al. utilized a language model to conduct concept-drift detection in junk mail filtering [5]. Lee et al. designed a rule-based model using the concept of decision trees to extract the concept-drift rules [12]. In addition, the concept of drift was often applied to classification and data stream [2, 6, 14, 16, 17, 20, 21].

Song et al. defined three types of concept-drift patterns, which could be focused in association-rule mining [18]. They are emerging patterns, unexpected changes, and added/perished patterns. Different types of concept-drift patterns indicate different meanings of concept drift for association rules. An evaluative function was then designed to calculate the degree of concept drift. If the degree between two rules is bigger than a predefined threshold, then a concept-drift pattern is generated. Hong et al. then generalized it and proposed fuzzy concept-drift patterns of association rules [10].

### 2.2 Fuzzy Data Mining

The fuzzy-set theory has been used in intelligent systems because of its simplicity and similarity to human reasoning [25]. The theory has been applied in fields such as manufacturing, engineering, diagnosis, and economics, among others. Many fuzzy data mining approaches have been proposed to solve real problems [8, 9, 11].

Srikant et al. proposed a mining method to handle quantitative transactions by partitioning the values of each attribute [19]. Hong et al. also proposed a fuzzy mining algorithm to mine fuzzy association rules from quantitative transaction data [8]. The fuzzy mining algorithm first used given membership functions to transform each quantitative value into a fuzzy set of linguistic terms and then used a fuzzy mining process to find fuzzy association rules. To handle the time-consuming mining task from a very large database, Fernandez-Basso et al. presented three Spark approaches to extract interesting fuzzy association rules from massive fuzzy data [3]. Based on the existing association rule mining algorithms of Apriori, Apriori-TID, and ECLAT, the map-reduced framework was utilized to speed up the data processing in mining fuzzy association rules. Besides, Zhang et al. proposed a differential evolution (DE) algorithm to extract important fuzzy association

rules and reduce spurious rules [26]. They also indicated that the rules extracted were better than those by non-evolutionary and genetic algorithms. They also gave a case study to show the DE-based approach was effective in practice.

### 2.3 Fuzzy C-means

FCM is a popular method for clustering. It follows the fuzzy-set theory and allows one piece of data to belong to two or more clusters [1]. It is frequently used in pattern recognition. The following objective function is minimized to get good clusters:

$$\sum_{j=1}^c \sum_{i=1}^n (u_{ij})^m \|c_j - x_i\|^2, \quad (1)$$

where  $m$  is an arbitrary real number greater than 1,  $x_i$  is the  $i$ -th data,  $c_j$  is the center of the  $j$ -th cluster,  $u_{ij}$  is the membership degree of  $x_i$  in the cluster  $j$ , and  $\|*\|$  is a norm expressing the similarity between a data and a center. The Euclidean distance is commonly used to calculate the norm.

FCM adopts an iterative process to minimize the above function. The process will stop when

$$\max_{i,j} \|u_{ij}^{k-1} - u_{ij}^k\| < \beta, \quad (2)$$

where  $\beta$  is a termination criterion between 0 and 1 and  $k$  is an iteration number. The algorithm of FCM is shown below.

#### The FCM algorithm

**Step1:** Initialize the  $U$  matrix,  $U^{(0)}$ , where  $U$  represents all the  $u_{ij}$  values.

**Step2:** At each  $k$ -th iteration, calculate the center of each cluster and the membership function of each data to each cluster by the following two formulas:

$$c_j = \frac{\sum_{i=1}^n (u_{ij})^m \times x_i}{\sum_{i=1}^n (u_{ij})^m}, \text{ and} \quad (3)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \frac{\|x_i - c_j\|^{\frac{2}{m-1}}}{\|x_i - c_k\|^{\frac{2}{m-1}}}}. \quad (4)$$

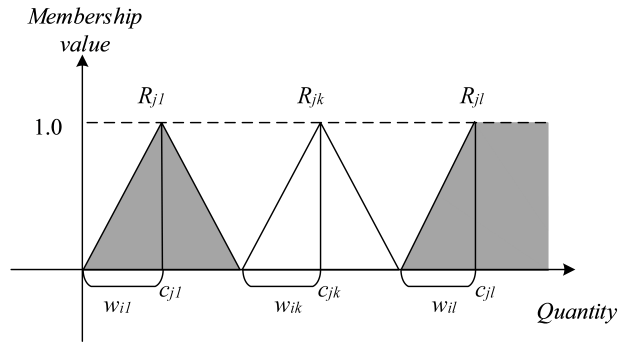
**Step3:** If formula (2) is reached, then stop; otherwise return to Step 2.

### 2.4 Fuzzy Data Mining

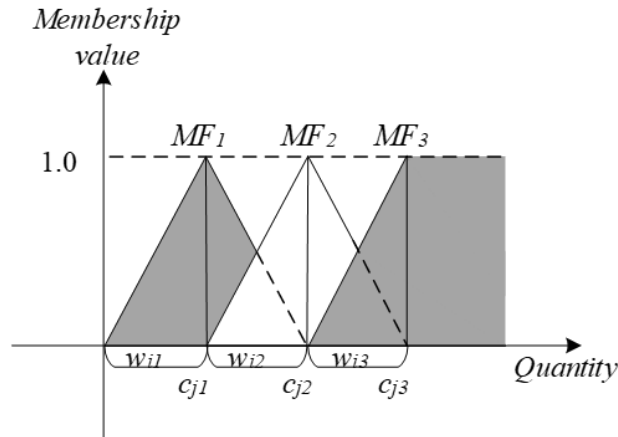
In this paper, we use the isosceles-triangle membership functions to represent the fuzzy regions [15] for simplicity. Isosceles-triangle membership functions are shown in Figure 1. The membership function of each fuzzy region  $R_{jk}$  is represented by a  $(c, w)$  pair, where  $c$  denotes the center abscissa and  $w$  represents half the span.

## 3 Concept Drift of Membership Functions

In this section, we present the concept-drift process of membership functions.



**Figure 1:** Representation of Membership functions.



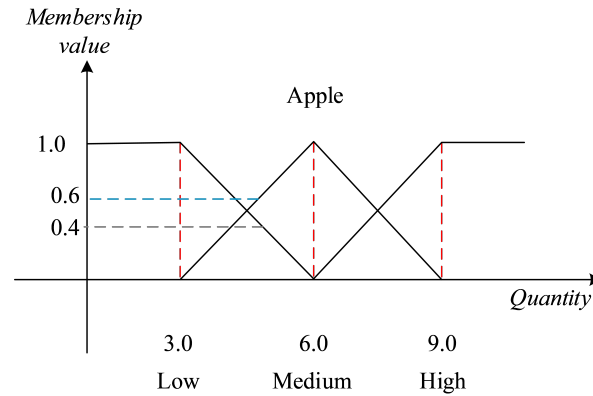
**Figure 2:** Membership functions of an item  $I_j$ .

### 3.1 Generating Membership Functions by Fuzzy $C$ -means

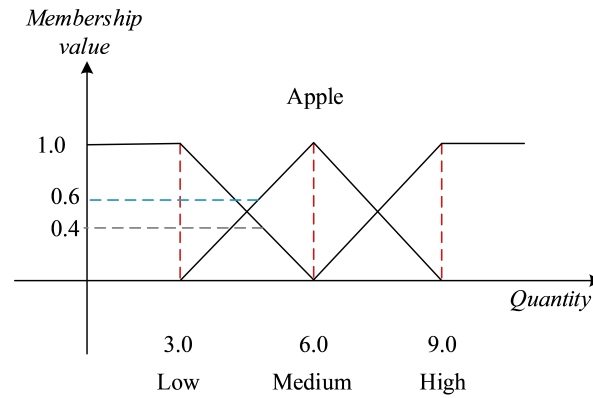
In this paper, we propose a simple approach based on FCM to generate a set of membership functions for an item. As mentioned above, a membership function is designed as an isosceles triangle and encoded as a pair of  $(c, w)$ . The peak of the triangle is located at  $c$ , and the distance between the peak and the left endpoint is  $w$ . If we need to generate  $n$  membership functions for an item, the proposed algorithm will obtain  $n$  cluster centers by using FCM. Each center obtained is the  $c$  value of the corresponding membership function. Then the span  $w$  is calculated as the distance between the location of the peak in this triangle with the previous one with the first one is the distance between the locations of the peak with 0. Figure 2 shows an example of three membership functions for an item  $I_j$ .

The membership functions play a critical role in converting commodity items into human semantics. Figure 3 shows a membership-function set for apples purchased in a transaction. It consists of three membership functions representing low, medium, and high, respectively, for each purchased amount. If we buy five apples, the low fuzzy value is 0.4, the medium fuzzy value is 0.6, and the high fuzzy value is 0.

Additionally, we may know the status of the concept from the membership functions. In Figure 3, the purchased amounts of three, six, and nine reach the membership value of 1 for the three membership functions, respectively. We can regard these amounts as the representative of the linguistic terms and observe the changes at different times.



**Figure 3:** An example of membership functions for purchased apple amounts.



**Figure 4:** Changed membership functions for the purchased number of apples.

### 3.2 Concept-drift Patterns of Membership Functions

We proposed three types of concept-drift patterns of membership functions. The first type is the change of the representative value (the center) of a linguistic term (membership function), the second type is the change of a linguistic-term span, and the third type is the change of the fuzzy support for a linguistic term. They are described below.

#### (A) The concept drift of the representative value for a linguistic term

Figure 3 shows the membership functions of the purchased number of apples derived in the last year, and Figure 4 shows those derived in this year. In the membership function for the linguistic term of low, the representative value, which is the center of the membership function, reduces from three to two. On the contrary, in the membership function for the linguistic term of high, the representative value increases from nine to ten. This represents the concepts of the low and the high linguistic terms have already changed.

The concept-drift degree of the representative value of a linguistic term, denoted  $cdLT$ , is thus shown below:

$$cdLT = \frac{|c_{ji}^{D'} - c_{ji}^D|}{w_{ji}^D}, \quad (5)$$

where  $D$  and  $D'$  are the initial and the final transaction databases at different times or different places,  $c_{ji}$

and  $w_{ji}$  are the center and the span values of the  $i$ -th linguistic term for the  $j$ -th commodity item. When the degree is larger than or equal to a given threshold, we may say it has a concept drift of a center.

As an alternative, we may also consider the average span as the denominator and set the formula below:

$$cdLT = \frac{|c_{ji}^{D'} - c_{ji}^D|}{\sum_{k=1}^N w_{jk}^D / N}, \quad (6)$$

where  $w_{jk}$  is the span value of the  $k$ -th linguistic term for the  $j$ -th commodity item, and  $N$  is the number of linguistic terms.

### (B) The concept drift of the span value for a linguistic term

The meaning of the span of a membership function is the coverage of a linguistic term on data. For example, although the representative value of the medium linguistic term is not changed in Figures 3 and 4, the span of the membership function from the modified database is larger than that from the original database. The concept-drift degree of the span value of a linguistic term, denoted  $cdMF$ , is thus designed below:

$$cdMF = \frac{|w_{ji}^{D'} - w_{ji}^D|}{(c_{jN}^D - c_{j1}^D) / (N - 1)}, \quad (7)$$

where  $D$  and  $D'$  are the initial and the final transaction databases at different times or different places,  $w_{ji}$  is the span values of the  $i$ -th linguistic term for the  $j$ -th commodity item,  $c_{j1}$  and  $c_{jN}$  are the first and the last center values of the  $j$ -th commodity item, and  $N$  is the number of linguistic terms. When the degree is larger than or equal to a given threshold, we may say it has a concept drift of a span.

### (C) The concept drift of the fuzzy support for a linguistic term

A concept drift in fuzzy support represents a group size changes for a membership function. We can use this value to measure concept drift. An example, may be the number of people that buy expensive mobile phones this year is greater than that in last year. We may use the following formula to evaluate the concept-drift degree of the fuzzy support value of a linguistic term, denoted  $cdSUP$ :

$$cdSUP = \frac{|sup_{ji}^{D'} - sup_{ji}^D|}{sup_{ji}^D}, \quad (8)$$

where  $D$  and  $D'$  are the initial and the final transaction databases at different times or different places,  $sup_{ji}$  is the span values of the  $i$ -th linguistic term for the  $j$ -th commodity item. When the degree is larger than or equal to a given threshold, we may say it has a concept drift of a fuzzy support. It can usually be used to represent the concept drift of customer purchase behaviour.

As an alternative, we may also consider the average support as the denominator and set the formula below:

$$cdSUP = \frac{|sup_{ji}^{D'} - sup_{ji}^D|}{\sum_{k=1}^N sup_{jk}^D / N}, \quad (9)$$

where  $N$  is the number of linguistic terms.

## 4 The Proposed Algorithm

In this section, the proposed approach that combines concept-drift and FCM is described. The algorithm is stated as follows.

### The algorithm for finding the concept drift of membership functions

**Input:**  $D$  and  $D'$ : databases;  $I$ : the number of items;  $S$ : concept-drift rule sets;  $M$ : the number of linguistic terms;  $\alpha, \beta, \gamma$ : the thresholds for judging the concept drift of centers, spans and fuzzy supports of membership functions, respectively.

**Output:** The concept-drift patterns of membership functions.

### Method:

**Step1:** Generate membership functions for each item from  $D$  and  $D'$  by the following substeps.

- (a) Use the FCM algorithm to find the center values of the  $N$  clusters of each item respectively for the two databases,  $D$  and  $D'$ .
- (b) Set the center points of these  $N$  clusters for each item as the centers of the membership functions.
- (c) Calculate the distances of all two neighbouring centers and set them as the spans of the membership functions.

**Step2:** Set the initial concept-drift pattern set  $S$  as  $\emptyset$ .

**Step3:** Find the three types of concept-drift patterns of membership functions of each item between  $D$  and  $D'$  by the following substeps.

- (a) Calculate the concept-drift degree ( $cdLT$ ) of the representative value of the linguistic term and compare it with the  $\alpha$  value. If the value of  $cdLT$  is larger than or equal to  $\alpha$ , then put the center drift pattern in  $S$ .
- (b) Calculate the concept-drift degree ( $cdMF$ ) of the span value of the linguistic term and compare it with the  $\beta$  value. If the value of  $cdMF$  is larger than or equal to  $\beta$ , then put the span drift pattern in  $S$ .
- (c) Calculate the concept-drift degree ( $cdSUP$ ) of the fuzzy support value of the linguistic term and compare it with the  $\gamma$  value. If the value of  $cdSUP$  is larger than or equal to  $\gamma$ , then put the support drift pattern in  $S$ .

**Step4:** After all the items are processed, output the concept-drift set  $S$ .

## 5 Experimental Results

In this section, we describe the experimental results of the concept drift of membership functions. We used a computer with an Intel Core i5 – 3230M 2.60GHz processor with four cores, four threads, and 12 GB RAM. The operating system used was Microsoft Windows 8.1 Pro and the programming language was .NET Framework 4.5.1 C# (C# Version 5.0).

A simulated retail dataset containing 1, 559 items and 21, 556 transactions was used in the experiments. In the dataset, the number of purchased items in transactions was first randomly generated, and the purchased items and their quantities in each transaction were then generated. Each transaction was also assigned a date and a location in one year.

**Table 1:** The numbers of concept-drift patterns of membership functions.

	Center Drift	Span Drift	Support Drift
Case 1	1	12	52
Case 2	9	39	234
Case 3	112	236	441
Case 4	40	112	274

The cluster size was set at 3, the fuzziness index value  $m$  in FCM was set at 2, the threshold values of  $\alpha$ ,  $\beta$ , and  $\gamma$  were all set at 1. We generate the following four cases to verify the concept drift of membership functions:

Case 1. Data from two different locations are selected to form the original and drifted databases.

Case 2. Data from the first and the second half of the dataset are selected to form the original and drifted databases.

Case 3. Data from two arbitrary months are selected to form the original and drifted databases.

Case 4. The whole data set is used as the original dataset and the data from an arbitrary month is selected to form the drifted database.

Table 1 shows the experimental concept-drift results of the proposed approach. In the experimental results, we can find the number of drift patterns at different times was larger than that at different locations. The short-term databases may contain more drifted patterns, so when we compared the membership functions from two short-term databases, more concept-drifts could be found. In the contrast, since the long-term databases tended to be stable, less concept-drifts will occur. As a result, for the simulated database, the comparison between short-term databases is more preferable. The experimental results with alternative formulas for concept-drift degrees are similar.

## 6 Conclusion and Future Work

In this paper, we have described a simple approach to test the concept drift of membership functions. We have proposed three types of concept drifts of membership functions and designed formulas to evaluate them. We have also implemented the approach based on fuzzy c-means (FCM) and the designed formulas. Experiments on the simulated retail dataset have also been made to show the effectiveness of the proposed approach. In particular, the proposed method can help shop managers understand customer behaviour drift, analyzed from membership functions, in different times and places. In the future, we will try to design more effective ways to decrease computing time and combine the proposed concept-drift patterns with fuzzy association rules. We will also conduct more experiments to verify the approach.

**Acknowledgements:** This research was supported by the Ministry of Science and Technology of the Republic of China under grant MOST 109-2221-E-390-015-MY3.

**Conflict of Interest:** The authors declare that there are no conflict of interest.

## References

- [1] J. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms, *Springer, New York*, (1981).



- [2] P. B. Dongre and L. G. Malik, A review on real time data stream classification and adapting to various concept drift scenarios, *IEEE International Advance Computing Conference, ITM University, 21-22 Feb., Gurgaon, India*, (2014), 533-537.
- [3] C. Fernandez-Basso, M. Ruiz and M. Martin-Bautista, Spark solutions for discovering fuzzy association rules in Big Data, *International Journal of Approximate Reasoning*, 137(07) (2021), 94-112.
- [4] H. Guo, H. Lia, Q. Rena and W. Wang, Concept drift type identification based on multi-sliding windows, *Information Sciences*, (585) (2022), 1-23
- [5] M. Z. Hayat, J. Basiri, L. Seyedhossein and A. Shakery, Content-based concept drift detection for Email spam filtering, *2010 5th International Symposium on Telecommunications, Iran Telecom Research Center, 4-6 Dec., Kish Island, Iran*, (2010), 531-536.
- [6] M. Z. Hayat, M. R. Hashemi, A DCT based approach for detecting novelty and concept drift in data streams, *2010 International Conference of Soft Computing and Pattern Recognition, Universite de Cergy-Pontoise, 7-10 Dec., Paris, France*, (2010), 373-378.
- [7] T. P. Hong, C. H. Chen, Y. L. Wu and Y. C. Lee, A GA-Based Fuzzy Mining Approach to Achieve a Trade-off Between Number of Rules and Suitability of Membership Functions, *Soft Computing*, 10(11) (2006), 10911101.
- [8] T. P. Hong, C. S. Kuo and S. C. Chi, Mining association rules from quantitative data, *Intelligent Data Analysis*, 3(5) (1999), 363-376.
- [9] T. P. Hong, K. Y. Lin and S. L. Wang, Fuzzy data mining for interesting generalized association rules, *Fuzzy Sets and Systems*, 138(09) (2003), 255-269.
- [10] T. P. Hong, J. M. T. Wu, Y. K. Li, and C. H. Chen, Generalizing Concept-Drift Patterns for Fuzzy Association Rules, *Journal of Network Intelligence*, 3(2) (2018), 126-137.
- [11] E. Hllermeier, Fuzzy sets in machine learning and data mining, *Applied Soft Computing*, 11(03) (2011), 1493-1505.
- [12] C. I. Lee, C. J. Tsai, J. H. Wu and W. P. Yang, A Decision Tree-Based Approach to Mining the Rules of Concept Drift, *Fourth International Conference on Fuzzy Systems and Knowledge Discovery Hainan University, 24-27 Aug., Haikou, China*, (2007), 639-643.
- [13] S. Mukkavilli and S. Shetty, Mining Concept Drifting Network Traffic in Cloud Computing Environments, *12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Carleton University, 13-16 May, Ottawa, ON, Canada, 05* (2012), 721-722.
- [14] E. Padmalatha, C. R. K. Reddy and B. P. Rani, Classification of Concept Drift Data Streams, *International Conference on Information Science Applications, Kyonggi University, 6-9 May, Seoul, Korea*, (2014), 1-5.
- [15] A. M. Parodi and P. Bonelli, A New Approach to Fuzzy Classifier Systems, *Proceedings of the 5th International Conference on Genetic Algorithms, University of Illinois Urbana-Champaign, 1 June, IL, USA*, (1993), 223-230.
- [16] P. D. Patil and P. Kulkarni, Adaptive Supervised Learning Model for Training Set Selection under Concept Drift Data Streams, *2013 International Conference on Cloud Ubiquitous Computing Emerging Technologies, Sri Venkateshwara College of Engineering, 15-16 Nov., Pune, India*, (2013) 36-41.

- [17] S. Shetty, S.K. Mukkavilli and L. H. Keel, An integrated machine learning and control theoretic model for mining concept-drifting data streams, *IEEE International Conference on Technologies for Homeland Security*, 15-17 Nov., Boston, USA, (2011), 75-80.
- [18] H. S. Song, J.K. Kim and S. Kim, Mining the change of customer behaviour in an Internet shopping mall, *Expert Systems with Applications*, 21(10) (2001), 157-168.
- [19] R. Srikant and R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables, *ACM Special Interest Group on Management of Data*, 25(2) (1996), 1-12.
- [20] J. Sun, H. Li and H. Adeli, Concept Drift-Oriented Adaptive and Dynamic Support Vector Machine Ensemble With Time Window in Corporate Financial Risk Prediction, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(4) (2013), 801-813.
- [21] B. Thuraisingham, Data mining for security applications: Mining concept-drifting data streams to detect peer to peer botnet traffic, *IEEE International Conference on Intelligence and Security Informatics*, 17-20 June, Taipei, Taiwan, (2008), 29-30.
- [22] A. Tsymbal, The Problem of Concept Drift: Definitions and Related Work, *Technical Report*, (2004).
- [23] D. H. Widyantoro and J. Yen, Relevant data expansion for learning concept drift from sparsely labeled data, *IEEE Transactions on Knowledge and Data Engineering*, 17(3) (2005), 401-412.
- [24] C. C. Yang and N.K. Bose, Generating fuzzy membership function with self-organizing feature map, *Pattern Recognition Letters*, 27(5) (2006), 356-365.
- [25] L. A. Zadeh, Fuzzy sets, *Information and Control*, 8(3) (1965), 338-353.
- [26] A. Zhang and W. Shi, Mining significant fuzzy association rules with differential evolution algorithm, *Applied Soft Computing*, 97 (2020), 105518.
- [27] X. Zheng, P. Li, X. Hu and K. Yu, Semi-supervised classification on data streams with recurring concept drift and concept evolution, *Knowledge-Based Systems*, 215(01) (2021), 106749.

**Tzung-Pei Hong**

Department of Computer Science and Engineering  
National Sun Yat-sen University  
Kaohsiung, Taiwan  
E-mail: tphong@nuk.edu.tw

**Chun-Hao Chen**


Department of Information and Finance Management  
National Taipei University of Technology  
Taipei, Taiwan  
E-mail: chchen@ntut.edu.tw

**Yan-Kang Li**

Department of Computer Science and Information Engineering  
National University of Kaohsiung  
Kaohsiung, Taiwan  
E-mail: m1025506@mail.nuk.edu.tw

**Min-Thai Wu**

College of Computer Science and Engineering  
Shandong University of Science and Technology  
Shandong, China  
E-mail: wmt@wmt35.idv.tw

©The Authors.  This is an open access article distributed under the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>) 