Dynamic optimisation based fuzzy association rule mining method

Noname manuscript No.
(will be inserted by the editor)

Dynamic optimisation Based Fuzzy Association Rule Mining Method

Hui ZHENG · Jing HE · Guangyan HUANG · Yanchun ZHANG · Hua WANG

Received: date / Accepted: date

Abstract Techniques of performance analysis, comprising of various metrics such as accuracy, efficiency and consuming time, have been conducted to evaluate the measures of properties and interestingness for the association rule mining method. Therefore, these metrics combined with different parameters (partitioning points, fuzzy sets) should be analysed thoroughly and balanced simultaneously to enhance the entire performance (effectiveness, accuracy and efficiency) for an algorithm. As a result, Most of the current algorithms face the pressure from the tradeoff of these metrics and parameters, which becomes even rougher when we employ it in different resources of data (discrete data, categorical data and continuous data). Specifically, serial data (i.e., sequences or transactions of floating point numbers), such as analysis of sensor streaming data, financial streaming data, medical streaming data and sentimental streaming data, are different from discrete variables, such as boolean data (e.g., sentiment: negative and positive represented as '0' and '1' separately) and categorical

Hui ZHENG
CAS Research Center on Fictitious Economy and Data Science,
University of Chinese Academy of Sciences, Beijing, 100190, China
Victoria University, Melbourne, Australia
and Fudan University, Shanghai, China
E-mail: hui.zheng2@live.vu.edu.au

Jing HE (corresponding author)
Institute of Information Technology, Nanjing University of Finance and Economics, College of Engineering and Science, Victoria University, Melbourne, Australia

Guangyan GUANG
School of Information Technology, Deakin University, Melbourne, Australia

Yanchun ZHANG
Centre for Applied Informatics, Victoria University, Melbourne, Australia
and Fudan University, Shanghai, China

Hua WANG
Centre for Applied Informatics, Victoria University, Melbourne, Australia

data (e.g., 'young age', 'middle age', 'old age'). The main difference is that serial data face sharp boundary's problem. That is, it is hard to decide the boundary values (i.e., the single points to partition data into different value groups), which is few to be solved in association rule mining methods. This paper aims to resolve the problem of sharp boundaries and balance multiple performances of our algorithm simultaneously by developing a novel dynamic optimisation (parameters and metrics) based fuzzy association rule mining (DOFARM) method. The proposed method can be applied in a wide range of classifying problems, such as the classification of sentiment strength (negative and positive). In our DOFARM method, instead of single partitioning points, we use a range of values to smoothly separate two consecutive partitions and develop a corresponding membership function to generate fuzzy sets for original data sets of physical and emotional diseases. Mainly, we design a dual compromise scheme: the first tradeoff balances better performance of out-putting association rules and more widely applicable fuzzy membership function while the second tradeoff reduces the time parameter as well as enhances the entire performance of our DOFARM method. The feasibility and accuracy of DOFARM we proposed have been certified theoretically and experimentally. Besides, we demonstrate the accuracy, effectiveness and efficiency for our DOFARM method by experiments according to both synthesis and real datasets.

Keywords Association Rule · Optimised Parameters · Multiple Objective Function · Data Mining

1 Introduction

Efficient analysis of serial data (i.e., sequences or transactions of floating point numbers) has become a crucial issue to be successfully resolved with the advancement of computing technology, such as data streams in financial, medical applications and physiological factors acquisition. Traditional classifiers can manage serial data and classify them into different groups conveniently. However, the hidden relationships in original data are also required to mine to provide further information, e.g., the possible product in a shopping process or the potential reason of type 2 diabetes. Association rule mining is therefore generally chosen for mining hidden relations and associations. The problem of the association rule mining method is that it concerns only non-continuous factors such as categorical sequence objects and customer transaction records and cannot handle continuous data quickly.

Suppose we have a constant feature: 'Age', a direct method is to divide this feature into intervals . When the number of intervals is fixed as three, we can use labels: 'young age', 'middle age' and 'old age' as the feature classes (crisp sets) after choosing the partitioning points. While, by using fuzzy theory [1] for the feature of 'Age', we can combine the three segments with membership functions by extending the boolean values 0 and 1 (respectively indicating absence and presence) to the continuous values from 0 to 1 ($[0, 1]$). Thus, the crisp transactions have been changed into fuzzy ones as shown in Figure 1. Specifically, crisp sets can only define whether a tuple contains an item, while in the fuzzy sets, we can define the degree of a tuple belonging to each interval. Still taking feature 'Age' as an example, we can generate

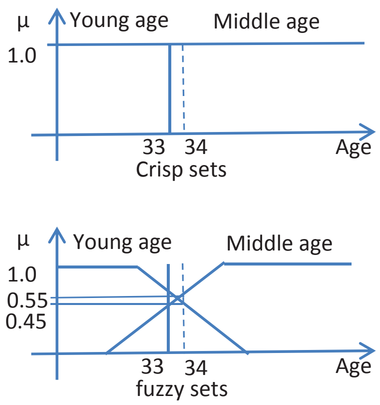

Fig. 1. An example of crisp sets and fuzzy sets.

three intervals $(0, 33]$, $(33, 67]$ and $(67, +\infty)$ with three classes 'young age', 'middle age' and 'old age' respectively. But it is non-reasonable to classify a person with 33 years old into the 'young age' class, an individual with 34 years old into the 'middle age' class. This problem is called sharp boundaries. In comparison, the fuzzy sets in Figure 1 can regard the 'Age': 34 as (young age, 0.45) (middle age, 0.55) instead of (young age, 0) (middle age, 1) in the crisp sets. Also, the feature of 'Age' can be applied for various purposes. Sometimes, we use 'Age' to judge the personal incomes; then we would like to change the partitioning points according to the modern working age and the modern retirement age. Sometimes, we distinguish 'Age' to measure the risk of heart disease or sentiment strength, in this situation, the feature 'Age' is better to be partitioned by using the changes of 'Age' rather than the absolute value. The characteristics of one person vary from gender, district, and country and all features may evolve. So all of these parameters, such as the partitioning points and fuzzy sets need to be improved and balanced simultaneously. When the continuous data are involved, it is not an easy task to extend the approaches introduced above.

As mentioned in paper [2], fuzzy logic was applied first to extend the association rule mining method with fuzzy sets of range, which keeps the advantage of numeric data with a membership value and diminishes the problem of the sharp or abnormal boundary in dividing the interval. Besides, a general model to discover association rules is proposed in work [3], which consists of the user-defined filter of certainty factors and the definition of very strong rules to generate interesting association rules. Apart from the paper [3], researchers have already presented some approaches to improve the method of fuzzy association rule mining. An assessment method to partition the data into different groups according to the features of data that are related to a given rule, that against the rule (the counterexamples) and that are irrelevant with the rule is developed in [4]. Another work in [5] introduces the novel measurements of quality by distinguishing the correlations of positive from the correlations of negative association rules; while extra measures (clustering, classifying, weighting and extracting membership function) are used to modify fuzzy association rule models [6], [7], [8] and [9].

Paper [10] proposes a classifying model called TME to distinguish social emotions of readers. Also, the generated topic indicators are utilised for the alleviation of overfitting problems. Additionally, the framework SenticRank of paper [11] aims to rank content-based sentiment and collaborative sentiment. Compared to it, this paper applies rule-based sentiment to further reveal to relationships between sentiments and individual features. As mentioned in papers [10], [12] and [13], most emotional or sentimental classification problems are solved by text mining. This paper will apply fuzzy association rule to partition sentiments into positive and negative, which means emotions benefit for well-being or harm to well-being.

As mentioned above, the method of fuzzy association rule discovering is not performed without its downside. The problems contain lacking the tradeoff scheme to select the most suitable partitioning points for association rules generating (while the continuous original data sets are transformed into fuzzy sets and corresponding membership values, the partitioning points are chosen as the points between any two adjacent fuzzy sets). The procedures of selecting partitioning points and calculating their membership values [3] and [4] are two essential processes of constructing fuzzy sets in building Fuzzy Association Rule Mining (abbreviation of FARM) model steps. Suppose $(0, 33]$, $(33, 67]$ and $(67, +\infty)$ of the feature 'Age' are three fuzzy sets with two partitioning points 33 and 67, which is not an accurate definition of middle age and it is inconsistent with general knowledge 44 and 59 in [14] or 40 and 60 in [15]. As the definition of middle age varies from domain, application and time - the particular algorithm is required to adjusting partitioning points regarding the accuracy improvement of distinguishing diabetes. However, there is not the task, on which researchers focus. What's more, any individual with 'Age' of 80 pay more attention with their emotional and physical well-being than a person with only 40 years old, but more and more kinds of diseases such as heart disease, diabetes and emotional disease are hitting on middle-aged people. So the relations with illness for the middle age (non-high value of feature) are ordinarily more useful than that related to old age (high value of feature). Also, the more related features (e.g., Age, systolic pressure, diastolic pressure, blood glucose) we consider, the more accurate result we can get. To be more specific, a person of middle age combined with other feature, i.e., systolic pressure of 130 mm HD, which is diagnosed as one of the two criteria of per hypertension (non-high value of element). In that case, a slight high in blood glucose will sharply increase the possibility of having diabetes problem than an old aged person with only a high value of 'Age'. Beyond this, the metrics of association-rules filtering and the parameters for the membership function smoothing still need to be improved and balanced simultaneously.

As these restraints of current FARM method, a generic method: Dynamic optimisation (parameters) based Fuzzy Association Rule Mining (DOFARM) is proposed, working with both continuous data and discrete data. It firstly offers a dual compromise scheme to balance the accuracy, effectiveness and efficiency of our algorithm simultaneously; Besides, the DOFARM method we proposed smoothes membership function of fuzzy sets and consequently reduces sharp boundary problems to a great extent. Moreover, our novel method which is based on the parameter selecting enhances the entire performances of fuzzy association rule mining by optimising parameters (partitioning points, fuzzy sets, the number of association rules) and metrics (support, confidence, certainty factor [3]). Fourthly, the efficiency - the most critical part of a method is improved almost two times by our DOFARM method when it skips some unnecessary steps with direction parameter selecting and reduces time-consuming of our DOFARM method. Therefore, combining these contributions of our DOFARM, we can say that it can finally achieve two interacting tradeoffs. To be more specific, it balances the effectiveness and accuracy (parameters: partitioning points and fuzzy sets) with multiple objective function scheme of the first tradeoff. Also, it adjusts the smoothly cognitive membership function and better performance of association rule mining (parameters and metrics of association rules) together, which is called dual compromise in this paper.

The rest of this paper is organised as follows. Section 2 describes the first tradeoff of our DOFARM method, which optimises three user-defined metrics to balance the accuracy and effectiveness of our algorithm simultaneously. In section 3, detailed algorithms of our second tradeoff are proposed through the interval partitioning, membership function constructing, parameters based metrics balancing and dual compromise mechanism proving. To further illustrate the DOFARM method and the dual compromise scheme we proposed, the computing processes and procedures are represented in section 4. The theorem, which demonstrates the universal applied dual compromise scheme is also illustrated in this chapter. The experimental performance evaluation of accuracy, effectiveness and efficiency of the proposed DOFARM method is studied in Section 5. Finally, conclusions are summarised in Section 6.

2 The First tradeoff: Balancing Different Metrics of Fuzzy Association Rules Simultaneously for Better Performance

Distinct from the classic fuzzy association rule mining method, our DOFARM method optimises frequent itemsets and association rules according to two additional tradeoff processings. It optimises the association rules of mining-output based on the previous frequent itemsets and the parameters based on selected metrics, which are used as metrics for optimisation functions. As a consequence, we can balance the effectiveness and accuracy using the proposed method (the second tradeoff). However, before that process, we should attempt to optimise the performance of our dynamic optimisation based fuzzy association rule mining, that is, balancing all of the metrics of fuzzy association rules: better-performed results and more interesting association rules (the first tradeoff).

2.1 A Multi-objective optimisation Scheme

In this subsection, we will introduce our scheme to optimise metrics of fuzzy association rules simultaneously which is based on Richardson Extrapolation and Gradient-based optimisation methods [16], [17], [18], [19], [20], [21]. A theorem is illustrated to indicate the correctness of our multi-objective optimisation process. Among all of the processes, the effective metrics should be listed and applied in our scheme firstly.

With the definition of multiple objectives, the problem that we are facing becomes optimising our metrics based objectives $\varphi_1, \varphi_3, \varphi_5, \varphi_{10}, \varphi_{n/2}$ simultaneously, by the procedure of selecting partitioning points according to the result of the direction from Algorithm 1.

Algorithm 1 Direction-computation Algorithm

Input: three thresholds: min_Supp, min_Conf and (min_CF; the user-defined number of objective function: N_objective (default value as 5); the total outputting number of association rules (n) a initialising set of partitioning points: $\mathcal{X}_0$; and the initialising gradients for every objective functions at the point of $\mathcal{X}_0$:

$$\mathbf{g}_1 = \nabla\varphi_1,\ \mathbf{g}_2 = \nabla\varphi_3,\ \mathbf{g}_3 = \nabla\varphi_5,$$
$$\mathbf{g}_4 = \nabla\varphi_{10},\ \mathbf{g}_5 = \nabla\varphi_{n/2},$$

Output: the chosen direction $\boldsymbol{\eta}$ (according to that direction all of the five objective functions can keep the condition of increasing in the limited area of the neighbourhood of the current point of $\mathbf{x}$).

1: At first: $\boldsymbol{\eta} \leftarrow \{0, \ldots, 0\}$.
2: **for** $i = 1, \ldots, N\_objective$ **do**
3: $\quad \boldsymbol{\alpha} \leftarrow \mathbf{g}_i$;
4: $\quad$ **for** $j = 1, \ldots, N\_objective$ **do**
5: $\quad\quad$ **if** $j \neq i$ and $\langle \boldsymbol{\alpha}, \mathbf{g}_j \rangle < 0$ **then**
6: $\quad\quad\quad \boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \frac{\langle \boldsymbol{\alpha}, \mathbf{g}_j \rangle}{\langle \mathbf{g}_j, \mathbf{g}_j \rangle} \mathbf{g}_j$;
7: $\quad\quad$ **end if**
8: $\quad$ **end for**
9: $\quad \boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \boldsymbol{\alpha}$;
10: **end for**
11: Return $\boldsymbol{\eta}$;

Theorem 1 *In every optimisation step, $\exists$ the direction $\boldsymbol{\eta}$ makes all of objective functions be optimised simultaneously.*

Proof Using the idea of Algorithm 1, the $\boldsymbol{\eta}$ can be calculated as

$$\boldsymbol{\eta} = \sum_{i=1}^{5} \sum_{j=1}^{5} (\mathbf{g}_i - \frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{\langle \mathbf{g}_j, \mathbf{g}_j \rangle} \mathbf{g}_j),$$

where $j \neq i$ and $\langle \mathbf{g}_i, \mathbf{g}_j \rangle < 0$.

With the above algorithm, we can sum up the Theorem 1. The parameter sets of partitioning points for the selected association rules are gradually optimised by the optimisation objective Algorithm 1. In the next section, we will consider about the optimisation procedure of the membership function and the corresponding fuzzy sets. Due to the necessity of keeping the priority of these metrics for multiple objective functions in our theorem, we should consider remaining this good performance of association rules in the second tradeoff.

2.2 The First tradeoff

The first tradeoff aims at balancing different metrics for the multiple objective functions mentioned above. For this purpose, Richardson Extrapolation formula and the steepest descent method are utilised and extended to multiple objective functions which can balance different metrics of fuzzy association rule simultaneously. The pseudo-code is shown in Algorithm 2. $Line$ 2 compute the parameters based objective functions, while $Line$ 3 calculate corresponding derivatives by Richardson Extrapolation method. Then, $Line$ 4 represents the processing of Algorithm 1 that computes the direction for increasing objective functions together. After that, we update the value of metrics φ through the selected path η and step size λ.

Algorithm 2 The First tradeoff: Balancing Multiple Metrics of Performance for Association Rule Mining

Input: initial (or previous) metrics φ, the maximum number of tradeoff rounds I;
Output: optimised parameters φ;
1: **for** $i = 0$ to I **do**
2: The processing of objective functions' computing;
3: The processing of derivatives' calculating [16] (Richardson Extrapolation approachis applied);
4: The processing of direction $\boldsymbol{\eta}$ selecting by Algorithm 1;
5: With the selected direction $\boldsymbol{\eta}$, we can update the objective functions to a larger one

$$\varphi \leftarrow \varphi + \lambda\boldsymbol{\eta},$$

where λ is a user-defined step size;
6: **end for**
7: Return φ;

3 The Second tradeoff: Balancing the Effectiveness and Accuracy of Our DOFARM Method

The first level of our dual compromise scheme aims at optimising all of the metrics for fuzzy association rules. While it has already optimised the preselecting metrics of association rule mining, maintaining this optimised performance of these preselecting metrics in the first level is becoming one of the basic tasks for our second tradeoff spontaneously. Also, the first tradeoff has not updated the partitioning points of the fuzzy-set membership functions with the parameter based metrics. Therefore this updating procedure should be considered in the second tradeoff. In the meantime, our dual compromise is still required to update the sets of frequent items and rules of generated from them according to the optimised partitioning points. The parameters related to the frequent itemsets are used to balance the number of elements in every fuzzy set of our method. Therefore, our DOFARM method will dynamically discover the optimised rules by the partitioning intervals and their frequent items of fuzzy sets, which can be used to analyse new coming data and supplied to decision-making efficiently [23], [24], [25], [26].

Algorithm 3 The Second tradeoff: Balancing Fuzzy sets and Partitioning Parameters

Input: the previous set of partitioning parameters $\mathcal{X}_0$ and the user-defined maximum number of rounds N_round_I (default as 5);

Output: optimised partitioning parameters $\mathcal{X}$; the optimised set of frequent item-sets F and the optimised set of Association Rules R;

1: Initialising $\mathcal{X}$ with the previous set of parameters $\mathcal{X}_0$;
2: Generate frequent item-sets F and association rules R, make sure R contains only strong rules;
3: **for** $i = 0$ to $\mathrm{N\_round\_Id}$ **do**
4:     **for** $\forall f \in F$ **do**
5:         Compute the value of $\mathrm{Supp}(f, \mathcal{X})$;
6:     **end for**
7:     **for** $\forall r \in R$ **do**
8:         Compute the value of $\mathrm{Supp}(r, \mathcal{X})$; Compute the value of $\mathrm{Conf}(r, \mathcal{X})$; Compute the value of $\mathrm{CF}(r, \mathcal{X})$;
9:     **end for**
10:     The processing of the objective functions' computing;
11:     The processing of the corresponding derivatives' calculating;
12:     The processing of the suitable directions' searching $\boldsymbol{\eta}$ according to Algorithm 1;
13:     Update the set of parameters with the searched direction $\boldsymbol{\eta}$ for larger value of our objective functions, update

$$\mathcal{X} \leftarrow \mathcal{X} + \lambda \boldsymbol{\eta},$$

where λ denotes step size;
14: **end for**
15: Return the current value of parameters $\mathcal{X}$, the present set of frequent fuzzy items F and the current set of association rules R;

The aim of our second tradeoff is to update the previous set of partitioning points generated from the first tradeoff. Also, the last set of association rules is applied when the current iteration is not the first one (Under-Optimised set of association rules is used in the first iteration). In our Algorithm 3, we can see that our the whole processing of the first tradeoff is shown as the $Lines$ 2 to 14 and the three thresholds are updated from $Line$ 4 to 9. After it, our multiple objectives optimisation procedure of the weighted parameter w is illustrated by the $Lines$ $10 - 13$, while we compute the proper direction of the Algorithm 1 through $Line$ 12.

4 The Dynamic optimisation based Fuzzy Association Rule Mining Method

In this section, we describe the fuzzy association rule mining method based on dynamic optimal parameters and metrics. The first subsection utilises an algorithm to further demonstrate the features of our dual compromise scheme and our DOFARM method. In the second subsection, the concrete steps of our DOFARM method are listed to interpret our process from the view of data processing further. As these processes and procedures displayed, we witness the operations of balancing the corresponding metrics (support, confidence and certainty factor) with the first tradeoff, and the methods of adjusting the current parameters (partitioning points, fuzzy sets) in the second tradeoff. Eventually, we conduct a global dual tradeoff between the predefined metrics and optimised parameters.

Apart from all of these details, another theorem based on the previous theorem proposed in the last section is also certified rigorously in this section. It not only demonstrates that the dual tradeoff enhances the performance of association rule mining theoretically but also illustrates a widely applied scheme to balance metrics of multiple functions and parameters related with high result-performing and low time-consuming simultaneously.

4.1 The dual compromise scheme

Our dual compromise scheme searches for the appropriate sets of association rules and frequent items through multi-aspect parameters, such as fuzzy sets and partitioning points improved by the second tradeoff. The pseudo-code is shown in Algorithm 4. The whole steps of our dual compromise scheme are introduced in $Lines\ 1-8$, while $Lines$ from 1 to 5 illustrate the processing of initialisation and the $Lines\ 6-8$ show the processing of how to optimise the set of partitioning points by the Algorithm 3.

Algorithm 4 The dual compromise scheme

Input: the original data set D; the threshold of support (min_Supp); the threshold of confidence (min_Conf); the user-defined maximum number of rounds in the second tradeoff N_round_J and the user defined maximum number of round in the first tradeoff N_round_I;

Output: the balanced set of frequent items F and the balanced set of association rules R;

1: $DP_1, \ldots, DP_4$ that is applied to distinguish different intervals is computed for a given continuous feature as described in paper [16];
2: **for** $\forall \mathcal{X}$ components x_0, x_1 in every continuous feature **do**
3: $\quad x_0 \leftarrow 0.5 * (DP_0 + DP_1)$;
4: $\quad x_1 \leftarrow 0.5 * (DP_2 + DP_3)$;
5: **end for**
6: **for** $j = 0$ to $\mathrm{N\_round\_J}$ **do**
7: $\quad$ The balance processing of optimising the set of fuzzy frequent items and partitioning parameters according to the Algorithm 3;
8: **end for**
9: Return the set of fuzzy frequent items F and the set of association rules of R by the $Line$ 7;

As the first tradeoff optimises the metrics by using multiple objective functions and the second tradeoff aims to balance the performance of fuzzy association rules and the partitioning points, the strong rules of the second tradeoff process will be different from that of the first tradeoff. To fulfil the dual tradeoff and it's optimising operations, the value of our multiple objective functions should be kept nondecreasing. Taking φ_{10}, which is one of the most popular metrics in association rule mining, as an instance, we have the Theorem 2, the other objective functions are just as the same situation as φ_{10}.

Theorem 2 *The value of objective function φ_{10} is non-decreasing during the dual tradeoff optimisation we proposed.*

Proof The optimisation we proposed consists of two levels of tradeoff. The second level of tradeoff reselect the association rules by redoing the frequent itemset discovering algorithm. The re-selection will either replace the original top 10 rules with ten better rules whenever it is possible or keep the original ten rules otherwise. So the dual tradeoff will either improve the value of φ_{10}, or maintain the value as it is. The first level of tradeoff perform a gradient-based Multi-Objective optimisation (we call it the first optimisation for convenience). The first tradeoff won't replace the top 10 rules; instead, it improves the quality of the top-ten rules since this quality is one of its objective function according to Theorem 1. So both the second level and the first level of our tradeoff ensure that the value of φ_{10} is nondecreasing.

In this way, our DOFARM method is proved as a generic measurement which can be widely used to balance multiple objective functions. This theorem means our objective function will be improved continuously both in the first and the second tradeoff. Precisely, in the first tradeoff, we optimise the preselecting metrics by using our objective functions, then we change the partitioning points to enhance the quality of the whole strong rule set in the second tradeoff. Afterwards the entire procedures of our dual compromise, we replace our entire strong rule set with better one. If improving the partitioning points of the fuzzy-sets will increase the number of rules above the given thresholds, then the dual compromise scheme we proposed will hopefully increase this number as well because the optimisation is performed with a set of objective functions that are related to the quality of the partitioning points. The experimental study will be illustrated to show the further achievements of our DOFARM method.

4.2 The Concrete Steps of the DOFARM Method We Proposed

Our DOFARM method is differing from the method of classic FARM concerning an additional dual tradeoff. It can optimise the set of frequent items and calculate the parameter based metrics, which are used as parameters for optimisation functions. The first tradeoff shown in Figure 2 can generate the optimised set of parameters for multiple objective functions. It optimises the performance of association rule entirely. Then, the second level of our tradeoff balances the partitioning points of the fuzzy-set membership functions based on optimal dynamic parameters. Eventually, our dual compromise scheme computes the set of frequent items and the set of association rules concerning the fuzzy sets optimised in the algorithm of the first tradeoff. The parameters are related to the frequent itemsets and used to balance the number of elements in every fuzzy set in our method. Therefore, our DOFARM method will dynamically discover the optimised set of association rules according to the continuously improving our multiple objective functions, which can be used to analyse new coming data and supplied to decision-making. All of the details will be presented in this section.

5 Experimental Study

There are four subsections in this section. The first subsection explains the corresponding methods, parameters and datasets. The second subsection lists the antecedents of strong rules and the results of partitioning points which indicate our

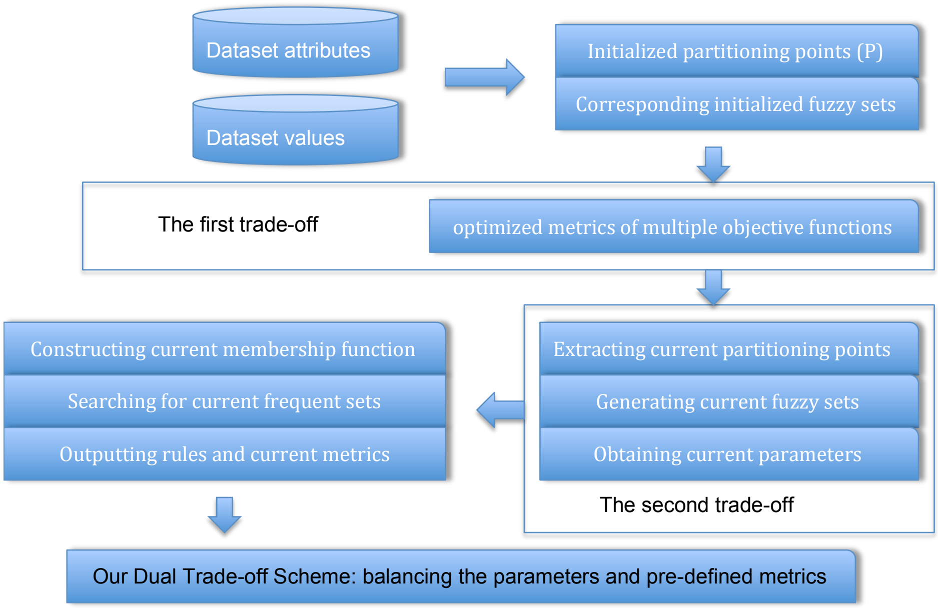

Fig. 2. Flowchart of Dynamic Optimal Parameter Based Fuzzy Association Rule Mining (DOFARM) method

rules have high accuracy and it follows the actual application as well. The third subsection illustrates the statistics of the three methods which further account for the effectiveness of our proposed DOFARM method. Finally, we compare our DOFARM with our previous work OFARM method [16] with data analysis of efficiency, as the method GFARM lose the general comparison conditions (details will be explained in this section). With all of the experimental studies, we can simply further represent the benefits of our DOFARM method, including the good performance of efficiency, effectiveness and accuracy expecting theoretical demonstrations in section 3.

5.1 Corresponding Methods and Experimental Datasets

In our experiment, the proposed DOFARM method is evaluated by comparing with GFARM method [3] and OFARM method of our previous work [16]. From the experimental descriptions among this section, we see our novel DOFARM method extends GFRAM and OFARM method to arbitrary parameters and metrics and improves it on accuracy, effectiveness and efficiency. The function of membership values for the GFARM, OFARM and our DOFARM method is already shown in the paper [16] and the strong rules are defined in [17]. A data set of “Pima Indians Diabetes” from UCI repository, is applied to display the outputting rules and compare the differences among the partitioning points of the three methods. The other data set coming from the Massachusetts General Hospital/Marquette Foundation (MGH/MF) Waveform

Database is applied to demonstrate the effectiveness and efficiency of our DOFARM method. The metric φ_{10} in section 2.1 is collected as one metric of effectiveness. The other metric of effectiveness is the number of the strong outputting rules. The user-defined maximum number of the second tradeoff algorithm and the number of the set of frequent fuzzy items are $N\_round\_J = 5$ and $q = 3$. The details of DOFARM we proposed are shown in Figure 2. The pruning method [22] of our experiments is applied to filter the set of association rules and prevent the huge amount of the number of rules. Following results in the three methods: GFARM, OFARM and our DOFARM will be shown as the average of five procedures, which is used to cut the randomness during our experiments.

The higher the value of thresholds are chosen, the better rules are generated, and then there will be a limited number of strong rules. So if the value of thresholds is set to be too high, the generated rules will normally be too narrow, while the value of thresholds is set to be too low, the quality of the generated rules will be too poor to be interesting. Thus, to manifest the exquisite adaptability of our DOFARM method, different thresholds of min_Supp and min_Conf are outputted and compared in our experiments. Therefore, we can prove the proposed DOFARM method according to a vast range of thresholds and then compare the differences.

5.2 Outputting of Strong Rules and Accuracy Comparisons

The set of association rules that are related to diabetes we discovered from “Pima Indians Diabetes” data set is represented in this subsection. All of the features and their IDs are described in the following items.

- 0: Number of times pregnant;
- 1: Plasma glucose concentration a 2 hours in an oral glucose tolerance test;
- 2: Diastolic blood pressure (mm Hg);
- 3: Triceps skin fold thickness (mm);
- 4: 2-Hour serum insulin (mu U/ml);
- 5: Body mass index (weight in kg/(height in m)^2);
- 6: Diabetes pedigree function;
- 7: Age (years);
- 8: Class variable (0 or 1).

The interesting and strong rules are defined and generated in this section. We firstly group continuous features from ‘0’ to ‘7’ into three sets of frequent fuzzy items. The leaving feature ‘8’ is a label of having diabetes or not (the value of ‘0’ is recognised as healthy people, and the value of ‘1’ represents the people who are suffering from diabetes). We only print the strong rules with their consequent $(8, 1)$ in our experiments, which denotes the $8 - th$ feature and its value is 1. So one of the interesting and strong rules can be shown as the form as $(4, 2)(7, 2) \rightarrow (8, 1)$. This outputted rule means when the second fuzzy set of the $4 - th$ feature and the second fuzzy set of the 7-th feature coincide; then the current individual can be indicated as diabetes. Our three thresholds are defined by $min\_Supp = 0.1, min\_Conf = 0.7$ and $min\_CF = 0.1$. All of the antecedents are shown separately without the

Table 1. Comparison of Interesting and Strong Rules in three methods.

Comparing Item	GFARM	OFARM	DOFARM
Antecedent $(4, 2)$	Containing	Containing	Containing
Antecedent $(7, 2)$	Containing	Containing	Containing
Antecedent $(1, 2)$	Containing	Containing	Containing
Antecedent $(3, 2)$	None	Containing	Containing
Antecedent $(5, *)$	None	$(5, 2)$	$(5, 1)$
Antecedent $(2, *)$	None	None	$(2, 1)$
Antecedent $(6, *)$	None	None	$(6, 1)$
Total Antecedent	3	5	7
Total Mid-Antecedent	0	0	3
Total rule	2	7	9

common consequent $(8, 1)$ in Table 1, where '*' means any possible value. Take $(4, 2)(7, 2) \rightarrow (8, 1)$ as instance, it will be divided into two antecedent $(4, 2)$ and $(7, 2)$.

Table 2. Partitioning points comparisons in three methods.

Model	Partitioning Points (Fuzzy Sets)
GFARM	$M_{L,0} = 1.5, M_{R,0} = 5.5;$
	$M_{L,1} = 102, M_{R,1} = 136;$
	$M_{L,2} = 66, M_{R,2} = 78;$
	$M_{L,3} = 25.5, M_{R,3} = 32.5283;$
	$M_{L,4} = 121.372, M_{R,4} = 168.519;$
	$M_{L,5} = 28.3, M_{R,5} = 35.75;$
	$M_{L,6} = 0.2615, M_{R,6} = 0.572;$
	$M_{L,7} = 25, M_{R,7} = 38;$
OFARM	$M_{L,0} = 1.8494, M_{R,0} = 6.9014;$
	$M_{L,1} = 95.0285, M_{R,1} = 125.031;$
	$M_{L,2} = 69.8885, M_{R,2} = 74.0415;$
	$M_{L,3} = 27.9882, M_{R,3} = 30.0862;$
	$M_{L,4} = 137.702, M_{R,4} = 158.103;$
	$M_{L,5} = 30.3525, M_{R,5} = 33.7087;$
	$M_{L,6} = 0.2198, M_{R,6} = 0.6890;$
	$M_{L,7} = 26.9063, M_{R,7} = 33.0771;$
DOFARM	$M_{L,0} = 1.9866, M_{R,0} = 6.9007;$
	$M_{L,1} = 108.966, M_{R,1} = 125.057;$
	$M_{L,2} = 69.902, M_{R,2} = 81.9074;$
	$M_{L,3} = 27.9887, M_{R,3} = 30.0819;$
	$M_{L,4} = 137.717, M_{R,4} = 163.632;$
	$M_{L,5} = 30.3597, M_{R,5} = 37.7868;$
	$M_{L,6} = 0.2197, M_{R,6} = 0.6890;$
	$M_{L,7} = 26.9798, M_{R,7} = 33.0393;$

According to the Table 1, we can observe that our proposed DOFARM method discovers seven different antecedents in all, while the OFARM gets five and GFARM has only three antecedents. Different from these common five antecedents with GFARM and OFARM, our DOFARM has two new antecedents (2,1) and (6,1), which means Diastolic blood pressure and Diabetes pedigree function have some relations with diabetes. The proposed DOFARM finds strong rules with more disease-related antecedents and more non-high antecedents. In real-world applications, the more amount of the features of the disease-related antecedents are, the more useful of the rule is. The DOFARM method we proposed, therefore, shows its first merit with two more disease-related antecedents. The second merit of our DOFARM method is that two rules associated with non-high value antecedents are discovered by our DOFARM method, while the methods of GFARM and OFARM find nothing. With general knowledge about association rule [22] and the interesting rule we defined (which is related to disease), the more interesting rules are filtered, better is the method. The GFARM and OFARM perform not well since they find less disease-related antecedents, less non-high value antecedents and less interesting rules. By contrast, our DOFARM generates a higher amount of interesting and strong rules, and it outputted rules seem to be more useful and productive in this light.

If the new continuous data is coming, we can use the same membership function and fuzzy sets defined by previous data to handle new data. Suppose there is an individual like Diastolic blood pressure of 80 mm Hg and 2-Hour serum insulin 164 mu $\mathrm{U/ml}$. Firstly, we can look up and find there are expressed as '2' and '4' respectively and transform them into fuzzy sets: Diastolic blood pressure (0, 0.8894, 0.1106) and 2-Hour serum insulin (0, 0.028, 0.972); Secondly, in Table 2, there are two related antecedents $(2, 1)$ and $(4, 2)$ and the rule $(2, 1)(4, 2) \rightarrow (8, 1)$ is found in our proposed DOFARM method; Thirdly, we can see the partitioning points in Table 1 as $M_{L,2} = 69.902, M_{R,2} = 81.9074$ and $M_{L,4} = 137.717, M_{R,4} = 163.632$; At last, we can see the individual have a high possibility of diabetes disease since the membership grades of $(2, 1)$ and $(4, 2)$ are high.

5.3 An example of segmental computing

Our algorithm of DOFARM can be widely applied in different applications, such as medicine, finance and affective and segmental computing. This subsection will illustrate an example of how our DOFARM applied in emotional and sentimental computing.

Emotions and sentiments have profound influences on medical treatments. In this paper, two sentiment strengths will be considered: positive (sentiment benefits for well-being) and negative (sentiment harms to well-being). For instance, people whose sentiment strengths are extremely positive would be active in treatments of controlling their unhealthy conditions. Patients with positive sentiment can enjoy their lives even if they are diagnosed with type 2 diabetes, coronary heart disease or cancers.

Subsection 5.2 shows the rules of diagnosing diabetes that can classify people into two groups: diabetes and nondiabetes. Our primary concern of classifying sentiment is the group of people who are diabetes, so we assume that the sentiment strength of nondiabetes will be extremely positive and then we entirely ignore this group in this subsection.

To be more simple and without loss of generality, we suppose only two attributes (body mass index and age) of the diabetes group that are related to sentiments. Then, applying our DOFARM on data of Table 3, the proposed method may generate rules related with sentiments as $BMI \leq 25$ and $Age \leq 33 \rightarrow sentiments : postive$ and $BMI \geq 30$ and $Age \geq 67 \rightarrow sentiments : negative$. Therefore, we can predict sentiments by the generated rules of our DOFARM.

Table 3. Sentimental data.

BMI	Age	Sentiments
60	23	postive
20	80	negative
...	...	...

5.4 Effectiveness Comparisons and Analysis

In this subsection, we use two metrics to evaluate our DOFARM's effectiveness accurately:

1. the number of rules;
2. the average value of top ten strong rules: φ_{10} in section 2.1, which combines values of min_Supp, min_Conf and min_CF, is used as metrics for the quality of rules.

The data set coming from the Massachusetts General Hospital / Marquette Foundation (MGH/MF) Waveform Database is applied to compare our proposed DOFARM method with GFARM and OFARM. For simplicity, we only take the record of mgh10 to assess our DOFARM method. The recording includes eight features: three ECG leads, arterial pressure, pulmonary arterial pressure, central venous pressure, respiratory impedance, and airway CO2 waveforms. We computed the average of five-time procedures to reduce the randomness of our experimental results.

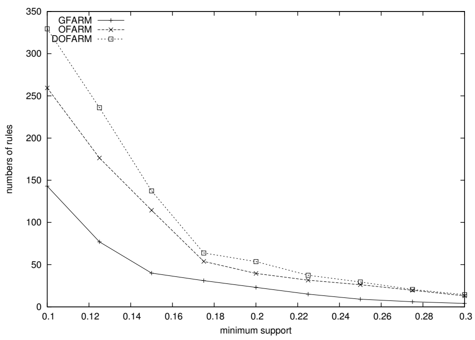

Fig. 3. Number of Rules comparison with different Minimum support

Among all these three methods: GFARM, OFARM and DOFARM, all of their number of rules show a downward trend with the growing min_Supp, which follows the property of min_Supp: the larger the min_Supp, the fewer association rules are filtered. However, there are still some differences in the changing process: the gap between OFARM and DOFARM is smaller than the difference between GFARM and OFARM at almost every point. That is to say, our DOFARM which is extended from OFARM not only inherits the benefits of OFARM but also exceeds the OFARM. Moreover, our DOFARM performs much better than other methods whether the original results are good or not (the min_Supp is small or large).

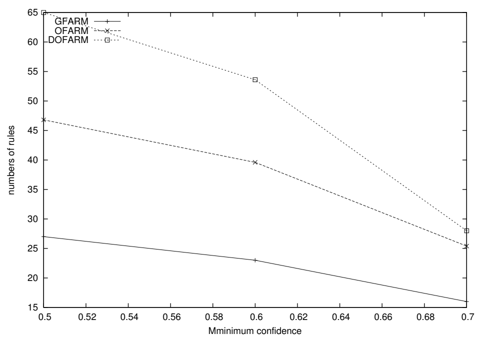

Fig. 4. Number of Rules comparison with different Minimum confidence

Combining the results in Figures 3 and 4, the number of rules for our method DOFARM is greater. The improvements of proposed DOFARM are more satisfactory when the original result is poor (with small min_Supp or small min_Conf) than the improvements with large min_Supp or large min_Conf. Then, the DOFARM method we proposed will get a much higher number of strong rules when the OFARM and GFARM are not good enough. Moreover, our DOFARM method can retain the benefits of OFARM method and can get a better result even if the results of OFARM are already pretty good.

The quality of fuzzy association rules is also used to verify the effectiveness of our DOFARM method. Taking the top ten rules as an example, the Figure 5 witnesses the quality of average of top ten rules decreasing according to the gradually increasing value of min_Supp with fixed min_Conf (0.6) and fixed min_CF (0.1). As we can manifest from the Figure 5, the φ_{10}, which means the average of quality (the minimum of support, confidence and certainty factor) of the top ten rules, drops when the min_Supp rises from 0.1 to 0.3. Among all these three methods, the DOFARM we proposed is always staying the highest column; the OFARM lies the column which is a little lower than the DOFARM column, while the GFARM is illustrated as the lowest. The difference of OFARM and our DOFARM in the histogram is still noticeable, and the column of our DOFARM shows its improvement at every different min_Supp in our experiments.

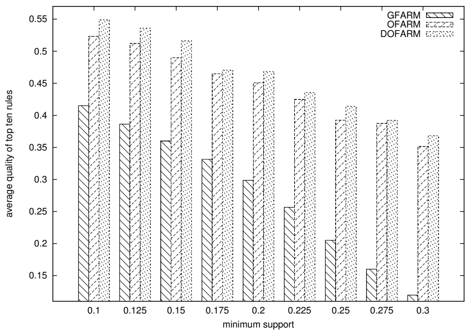

Fig. 5. Top Ten rules quantity comparison with different Minimum support

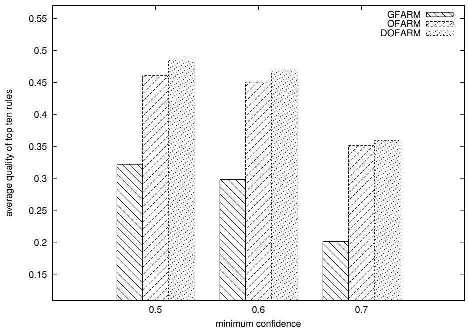

Fig. 6. Top Ten rules quantity comparison with different Minimum confidence

The situation of quality φ_{10} with the increasing value of min_Supp is just similar to the case in increased min_Conf with fixed min_Supp (0.2) and min_CF (0.1), which is indicated in Figure 6. The column of our DOFARM is still higher than the other two columns of GFARM and OFARM. So we can say our DOFARM can generate more suitable rule sets than other compared methods.

With Figure 5 and 6, our DOFARM method demonstrates greater effectiveness comparing with the GFARM and OFARM method. Then, the proposed DOFARM method outperforms the other two methods concerning both the quantity of outputted rules and the quality of interesting rules. To be more specific, optimising the set of partitioning parameters enhances the amount of our outputted rules; while the selected parameters of the functions of objectives increase the quality of interesting rules with the thresholds of min_Supp and min_Conf.

5.5 Efficiency Comparisons and Analysis

In this section, we consider estimating the efficiency of our DOFARM method. To certify the performance of our DOFARM method, we compare our method with the previous work OFARM method. As GFARM method runs once only and OFARM, DOFARM method runs several times to balance the different metrics of fuzzy association rule mining process, GFARM method lose the necessity of comparison while parameter time is related.

Following the general rule, we use the program running time t and the two effectiveness metrics: the number of strong rules N and the average of the quantity of best ten rules φ_{10} together, then generate the formula for efficiency as follows:

$$Efficiency = \frac{N \cdot \varphi_{10}}{t}.$$

For example, when min_Supp = 0.3 and min_Conf = 0.7, we can compute $Efficiency$ of OFARM and DOFARM by $Efficiency^{OFARM}_{(0.3,0.7)}$ and $Efficiency^{DOFARM}_{(0.3,0.7)}$ as follows:

$$Efficiency^{OFARM}_{(0.3,0.7)} = \frac{25.4 \times 0.3766452}{21.18} = 0.45168971 \approx 0.4517.$$

$$Efficiency^{DOFARM}_{(0.3,0.7)} = \frac{28 \times 0.3866068}{11.0388} = 0.980631083 \approx 0.9807.$$

As this formula illustrated, we sort out the $Efficiency$ with different min_Supp from 0.1 to 0.3 and fixed min_Conf (0.6) in Table 4.

Table 4. Efficiency comparison with different Min-support.

Min-support	0.1	0.125	0.15	0.175	0.2
OFARM	2.7118	1.6863	1.8614	0.7225	0.9366
DOFARM	4.6485	3.2722	3.8347	1.8567	3.9215
Min-support	0.225	0.25	0.275	0.3	Average
OFARM	0.5047	0.1918	0.1548	0.1277	0.9886
DOFARM	0.7902	0.3067	0.2697	0.1384	2.1154

As it is described in Table 4, we can witness the good performance of the DOFARM we presented, as the $Efficiency$ of our DOFARM is always larger than the $Efficiency$ of OFARM method. Also, the differences between our DOFARM and the OFARM method show decreasing trends with more and more strict threshold (min_Supp) from 0.1 to 0.2. During the period of changing the value of min_Supp, the gap shrinks marginally from 0.225 to 0.3. More particularly, in Table 4 there are nearly four times $Efficiency$ of our DOFARM at $min\_Supp = 0.2$ than the $Efficiency$ of OFARM, which demonstrate the DOFARM we proposed in this paper is almost two times of $Efficiency$ as its counterpart.

The $Efficiency$ with different min_Confs and fixed $min\_Supp = 0.2$ are illustrated in Table 5, which also describe better performance of efficiency of our DOFARM method than that of the OFARM method. From Table 5, we can further demonstrate the $Efficiency$ of the DOFARM method we proposed is much higher than the $Efficiency$ of OFARM method. In summary, we can conclude that our DOFARM performs much better than other two methods: GFARM and OFARM in

Table 5. Efficiency comparison with different Min-confidence.

Min-confidence	0.5	0.6	0.7
OFARM	0.7124	0.9366	0.4517
DOFARM	1.9896	1.4543	0.9806

both effectiveness and accuracy metrics. So the DOFARM method is demonstrated as the better one to generate larger amount of quantity of strong rules and better quality efficient rules.

6 Conclusion

A dynamic optimisation fuzzy-association-rule mining method has been proposed according to the definition of the dual compromise measurement. We have shown that the balancing procedures of the parameter-based-metrics make the proposed method easy to formulate and valid to balance parameters and metrics simultaneously for continuous data. In the algorithm of our dual compromise, the set of fuzzy association rules and the set of frequent items are optimised by the selected set of partitioning parameters. After outputting association rule of the three methods GFARM, OFARM and our DOFARM, the accuracy of the rule sets has been certified. The experiment also demonstrates that our DOFARM method is capable of balancing the parameters of the quality of interesting rules and the quantity of outputted rules; that is, the effectiveness of our DOFARM method exceeds the other two approaches. Also, after comparing with OFARM method, the efficiency of the DOFARM we proposed is almost two times of its counterpart, as the consuming time of our method has been reduced to half as OFARM method averagely. In conclusion, our DOFARM method outperforms its peers - GFARM and OFARM in accuracy, effectiveness and efficiency. Furthermore, the results of our experiments for gradual changes of min_Supp and min_Conf show the stability and robustness of the DOFARM method we proposed. In this paper, problems of both diabetes and sentiment strength employ our DOFARM method to accomplish their solutions of classification.

Acknowledgements This work is supported in part by The ARC Discovery Early Career Research Award (DE130100911), The ARC Discovery Project (DP130101327), The ARC Linkage Project (LP100200682), The NSFC funding(61332013), The International Science and Technology Cooperation Projects(No.2016D10008, 2013DFG12810, 2013C24027), The Municipal Natural Science Foundation of Ningbo(No.2015A610119), The Natural Science Foundation of of Zhejiang Province (No. Y16F020002).

References

1. Q. Zhou, P. Shi, S. Xu, H. Li, "Adaptive Output Feedback Control for Nonlinear Time-Delay Systems by Fuzzy Approximation Approach," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 2, pp. 301-313, 2013.
2. J.H. Lee and H.L. Kwang, "An extension of association rules using fuzzy sets," *In Proceedings of the Seventh IFSA World Congress*, pp. 399-402, 1997.

3. M. Delgado, N. Marłn and D. Snchez, “Fuzzy association rules: general model and applications,” *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 2, pp. 214-225, 2003.
4. X.Z. Wang, RAR. Ashfaq, A.M. Fu, “Fuzziness based sample categorization for classifier performance improvement”, *Journal of Intelligent and Fuzzy Systems*, vol. 29, no. 3, pp. 1185-1196, 2015.
5. M. De Cock, C. Cornelis, E. E. Kerre, “ Fuzzy Association Rules: a Two-Sided Approach,” *In Proceedings of Internal Conference on Fuzzy Information Processing-Theories and Applications*, pp. 385-390, 2003.
6. C.L. Chen, FSC. Tseng, T. Liang, “An integration of WordNet and fuzzy association rule mining for multi-label document clustering,” *Data and Knowledge Engineering*, vol. 69, pp. 1208-1226, 2010.
7. J. Alcalá-Fdez, R. Alcalá, and F. Herrera,“A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning,” *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857-872, 2011.
8. V.V. Rao, E. Rambabu, and G. Sriramganesh, “Effective Association rule mining using Fuzzy Apriori and Weighted Fuzzy Apriori,” *IJECCE*, vol. 3, no. 3, pp. 381-386, 2012.
9. J. Alcalá-Fdez , R. Alcalá, M.J. Gacto, and F. Herrera, “Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms,” *Fuzzy Sets and Systems*, vol. 160, No. 7, pp. 905-921, 2009.
10. Y. Rao, H. Xie , J. Li and F. Jin , F. L. Wang, and Q. Li , “Social emotion classification of short text via topic-level maximum entropy model,” *Information and Management*, vol. 53, No. 8, pp. 978-986, 2016.
11. H. Xie , X. Li , T. Wang , R. Lau , T. L. Wong, L. Chen, and Q. Li , “ Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy,” *Information Processing and Management*, vol. 52, No. 1, pp. 61-72, 2016.
12. A. Dridi, and D. R. Recupero , “Leveraging semantics for sentiment polarity detection in social media,” *International Journal of Machine Learning and Cybernetics*, vol. 1, No. 11, 2017.
13. W. Wang , G. Tan, and H. Wang , “Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews,” *International Journal of Machine Learning and Cybernetics*, vol. 8, No. 3, pp. 1053-1070, 2017.
14. O. B. Omar ,C. Boschi-Pinto, and A. D. Lopez, “AGE STANDARDIZATION OF RATES: A NEW WHO STANDARD,” *World Health Organization*, vol. 13, no. 2, pp. 167-192, 2001.
15. X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, “News impact on stock price return via sentiment analysis”’, *Knowledge-Based Systems*, vol. 69, pp. 14-23, 2014.
16. H. Zheng, J. He, G.Y. Huang, and Y.C. Zhang, “optimised Fuzzy Association Rule Mining for Quantitative Data,” *In Proceedings of 2014 IEEE International Conference on Fuzzy Systems*, pp. 396-403, 2014.
17. M. Delgado, D. SáNchez, M. J. MartıN-Bautista, and M.A. Vila, “Mining association rules with improved semantics in medical databases.” *Artificial Intelligence in Medicine*, vol. 21, no. 1, pp. 241-245, 2001.
18. J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang , “A probabilistic method for emerging topic tracking in microblog stream.” *World Wide Web*, vol. 20, no. 2, pp. 325-350, 2017.
19. J. Ma, L. Sun, H. Wang, Y. Zhang, and U. Aickelin, “Supervised anomaly detection in uncertain pseudoperiodic data streams.” *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 1, pp. 1-4, 2016.
20. H. Wang, Y. Zhang, and J. Cao, “Effective collaboration with information sharing in virtual universities.” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 6, pp. 840-853, 2009.
21. H. Wang, J. Cao, and Y. Zhang, “A flexible payment scheme and its role-based access control.” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 425-436, 2005.
22. J. Han, M. Kamber, and J. Pei,” *Data mining: concepts and techniques*, Morgan kaufmann, pp. 229-242, 2006.
23. J. Zhang, X. Tao, and H. Wang,” *Outlier detection from large distributed databases*, World Wide Web, vol. 17, no. 4, pp. 539-568, 2014.
24. M.E. Kabir, H. Wang, and E. Bertino,” *Efficient systematic clustering method for k-anonymization*, Acta Informatica, vol. 48, no. 1, pp. 51-66, 2011.
25. F. Khalil, J. Li, and H. Wang,” *An integrated model for next page access prediction*, International Journal of Knowledge and Web Intelligence, vol. 1, no. 1, pp. 48-80, 2009.
26. X. Sun, H. Wang, J. Li, and J. Pei,” *Publishing anonymous survey rating data*, Data Mining and Knowledge Discovery, vol. 23, no. 3, pp. 379-406, 2011.