IRBM 42 (2021) 198–206

Contents lists available at ScienceDirect

IRBM

www.elsevier.com/locate/irbm

Elsevier Masson France

EM|consulte

www.em-consulte.com

Original Article

Detection of Breast Cancer Based on Fuzzy Frequent Itemsets Mining

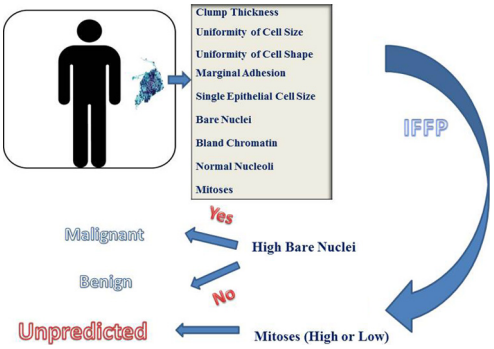
F. Ramesh Dhanaseelan, M. Jeya Sutha *

Dept. of Computer Applications, St. Xavier's Catholic College of Engineering, Nagercoil – 629 003, Tamil Nadu, India

HIGHLIGHTS

- Investigates the core factors that contribute to breast cancer.
- A new fuzzy Apriori based algorithm is introduced to analyze the biological dataset.
- The algorithm detects whether the person belongs to malignant or benign.
- Proposed algorithm's efficiency is proved.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 July 2019
Received in revised form 13 January 2020
Accepted 6 May 2020
Available online 19 May 2020

Keywords:
Data mining
Fuzzy frequent itemsets
Breast cancer
Fuzzy logic
Crisp set

ABSTRACT

Background: Breast cancer, a type of malignant tumor, affects women more than men. About one third of women with breast cancer die of this disease. Hence, it is imperative to find a tool for the proper identification and early treatment of breast cancer. Unlike the conventional data mining algorithms, fuzzy logic based approaches help in the mining of association rules from quantitative transactions.
Methods: In this study a novel fuzzy methodology IFFP (Improved Fuzzy Frequent Pattern Mining), based on a fuzzy association rule mining for biological knowledge extraction, is introduced to analyze the dataset in order to find the core factors that cause breast cancer. This method consists of two phases. During the first phase, fuzzy frequent itemsets are mined using the proposed algorithm IFFP. Fuzzy association rules are formed during the second phase, indicating whether a person belongs to benign or malignant. This algorithm is applied on WBCD (Wisconsin Breast Cancer Database) to detect the presence of breast cancer.
Results: It is determined that the factor, Mitoses has low range of values on both malignant and benign and hence it does not contribute to the detection of breast cancer. On the other hand, the high range of Bare Nuclei shows more chances for the presence of breast cancer.
Conclusion: Experimental evaluations on real datasets show that our proposed method outperforms recently proposed state-of-the-art algorithms in terms of runtime and memory usage.

1. Introduction

Breast cancer is one of the diseases that mostly affect women and is the second leading cause of cancer death [1]. It is the most dangerous disease because about one third of women with breast cancer die of the disease, although it is curable when detected early [2]. Women over the age of 50 are mostly affected by this

* Corresponding author.
E-mail addresses: message_to_ramesh@yahoo.com (F. Ramesh Dhanaseelan), jayasuthaus@rediffmail.com (M. Jeya Sutha).

https://doi.org/10.1016/j.irbm.2020.05.002
1959-0318/

disease. Every year about 48,000 cases occur in the UK [3]. Around one in nine women is affected by this disease at some stage in their life. It can easily be cured if diagnosed at an early stage. Therefore it is necessary for the proper identification and early treatment of the disease.

Mammography is one of the frequently used methods to detect breast cancer [4]. The value of mammography can identify breast abnormalities with 85-90% accuracy [5]. Fine needle aspiration cytology (FNAC) is also widely used in the diagnosis of breast cancer. Still, its average correct identification rate is only 90% [6]. Hence, it is necessary to develop better identification method to diagnose the breast cancer. Several researchers have employed statistical techniques and artificial intelligence techniques to predict breast cancer [7,8]. The main objective of these identification techniques is to identify whether the person belongs to either a benign group that does not have breast cancer or a malignant group that has a strong evidence of having breast cancer. So, the diagnosis of breast cancer problems is more general and widely discussed classification problem [9–12].

Previous studies refer to a number of techniques to diagnose breast cancer pattern. Neural network, a classification method based on which many algorithms [4,13–17] have been developed for diagnosing breast cancer. Artificial neural networks and multivariate adaptive regression splines approach [4], association rules and neural network approach [13], radial basis function neural network classification technique [18], Genetic algorithm based approach [19] and support vector machines (SVM) [20–24] are some of the techniques used in breast cancer detection. A data separation/classification method called isotonic separation technique [25] is one of the methods followed in predicting breast cancer. In Salama et al., [26] different classifiers like multilayer perceptron neural network, combined neural network, probabilistic neural network, recurrent neural network and SVM were analyzed for classification accuracies of breast cancer detection. Lu et al. [27] proposed an automated computer aided diagnosis framework which consists of ensemble under-sampling (EUS) for imbalanced data processing, the relief algorithm for feature selection, the subspace method for providing data diversity, and Adaboost for improving the performance of base classifiers. They extracted morphological, various textures, and Gabor features for magnetic resonance imaging (MRI). Wang et al. [28] proposed an SVM-based ensemble learning algorithm to reduce the diagnosis variance and increase diagnosis accuracy. Sivakumar et al. [29] developed an algorithm for breast cancer diagnosis based on Supervised Learning in Quest (SLIQ) and Decision Tree algorithms. Peng at al. [30] proposed an automated breast cancer diagnosis algorithm which organically integrates artificial immune with semi-supervised learning. Jafari-Marandi et al. [31] presented a data and decision analytic method that employs both supervised and unsupervised learning powers of ANNs to optimize breast cancer diagnosis with regard to decision-making goals. Alwidian et al. [32] developed a new technique based on a weighted method to select more useful association rules and a statistical measure for pruning rules for breast cancer disease.

Data mining plays the main role in the detection of knowledge from medical data repositories that could benefit medical diagnosis and for disease prevention [33]. Most of conventional data mining algorithms find the relation among transactions with binary values. However, transactions with quantitative values are commonly seen in real world applications. Fuzzy logic based approaches [34–39] take the major role in mining of association rules from quantitative transactions.

Several fuzzy mining methods have been proposed for finding interesting linguistic association rules from transaction data with quantitative values. Methods can be divided into two types: level-wise [40–44] algorithms and pattern-growth [34,45–48] algorithms. Level-wise approach generates patterns containing 1 items, then 2 items, 3 items, etc. It repeatedly scans the database to count the support of each pattern. On the other hand, pattern-growth algorithms utilize a depth-first search instead of a breadth first search and it only considers patterns actually existing in the database. Chan and Au proposed an algorithm called the F-APACS, a level-wise approach for mining fuzzy association rules [40] in which quantitative attribute values are transformed into linguistic terms and then the adjusted difference analysis is used to find interesting associations among attributes. Hong et al. proposed the fuzzy mining algorithms for mining fuzzy association rules from quantitative transaction data [41–43]. Kuok et al. proposed a fuzzy mining method for handling numerical data in databases to derive fuzzy association rules [44]. There are several methods have been implemented for fuzzy data mining based on pattern-growth approach. Papadimitriou and Mavroudi proposed the fuzzy frequent-pattern tree mining algorithm (FFPT) for finding fuzzy association rules [48]. It considers only two linguistic terms to construct the tree structures. The transformed membership values of linguistic terms are then checked against the user specified minimum support threshold.

Lin et al. proposed the algorithms fuzzy FP-tree [45] and compressed fuzzy FP-tree (CFFP-tree) [46] for efficiently mining fuzzy frequent itemsets from a quantitative dataset. Single linguistic term with the maximum membership value among transformed linguistic terms of an item is handled by these algorithms. Hong et al. extend the fuzzy FP-tree into multiple fuzzy-term FP (MFFP) tree [47] to mine all fuzzy frequent itemsets, instead of the representative linguistic terms from a set of quantitative transaction. Lin et al. proposed the compressed multiple fuzzy frequent pattern (CMFFP)-tree [34] algorithm to mine complete multiple fuzzy frequent itemsets. It is designed to keep not only the linguistic term with maximum membership value but also the other frequent linguistic terms for mining the completely fuzzy frequent itemsets. In this algorithm, the multiple frequent linguistic terms are sorted in descending order of their occurrence frequencies to build the CMFFP-tree structure. Every node in the CMFFP-tree uses an additional array to maintain the membership values of its prefix path by intersection operation.

In this study, a novel fuzzy methodology, IFFP (Improved Fuzzy Frequent Pattern Mining) has been introduced, which is based on a fuzzy association rule mining for biological knowledge extraction. The method consists of two phases. During the first phase fuzzy frequent itemsets are mined using the proposed algorithm IFFP. In the second phase, fuzzy association rules (ARs) are formed which indicate whether the person belongs to benign category (not dangerous) or malignant category (dangerous to health).

The remaining of this paper is organized as follows. Section 2 illustrates the datasets used; Section 3 presents the preliminaries related to the proposed algorithm; Section 4 introduces the algorithm; Section 5 describes the application of algorithm on the dataset; Section 6 presents the experimental results and finally Section 7 concludes the paper with a summary of the findings.

2. Wisconsin breast cancer database

Breast cancer is a type of malignant tumor that develops from breast cells. The reason for the cell to become cancerous may be that something damages or alters certain genes in the cell which in turn makes the cells abnormal and make them multiply beyond control. It occurs both in men and women, however male breast cancer is rare. Although scientists list out some of the risk factors (family history, obesity, infertility, ageing, menstrual periods, genetic risk factors) that make a women's chances of developing breast cancer [13], they do not know exactly how these risk factors cause cells to become cancerous. Their researches are underway to

Table 1
WBCD description of attributes.

Attribute	Domain	Mean	Standard deviation
Sample Code Number	id number	–	–
Clump Thickness	1 – 10	4.42	2.82
Uniformity of Cell Size	1 – 10	3.13	3.05
Uniformity of Cell Shape	1 – 10	3.20	2.97
Marginal Adhesion	1 – 10	2.80	2.86
Single Epithelial Cell Size	1 – 10	3.21	2.21
Bare Nuclei	1 – 10	3.46	3.64
Bland Chromatin	1 – 10	3.43	3.64
Normal Nucleoli	1 – 10	2.87	3.05
Mitoses	1 – 10	1.59	1.71
Class	2 for benign, 4 for malignant	–	–

Table 2
Example dataset.

TID	Items
1	(A:9) (B:5) (C:2)
2	(A:10) (B:2) (C:6)
3	(A:8) (B:4) (C:3)
4	(A:10) (B:5) (C:2)

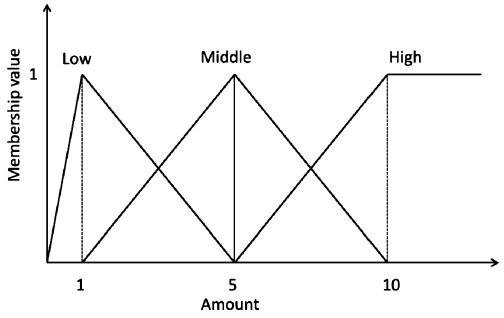


Fig. 1. Fuzzy Membership function.

learn more and make great progress in understanding how certain DNA changes can cause normal breast cells to become cancerous [49].

In this paper, the Wisconsin breast cancer database (WBCD) was used and analyzed [50]. The dataset has been collected by Dr. William H. Wolberg (1989-1991) at the university of Wisconsin-Madison Hospitals. Features are determined from a digitized image of a fine needle aspirate (FNA) of a breast cancer. Computed features describe the characteristics of the cell nuclei present in the image. Dataset of WBCD shown in Table 1 has 11 attributes including the class attribute. Every attribute except the code number and class has the domain ranges from 1 to 10 where 1 is the normal condition and 10 is the most critical condition. The database has two classes called as benign and malignant. There are 699 instances. Among them 241 records (34.5%) belong to malignant category and 458 records (65.5%) belong to benign category.

3. Preliminaries

Let $D = \{T_1, T_2 \ldots T_N\}$ be a quantitative database with M items, $I = \{i_1, i_2, \ldots i_M\}$ where T_i is the ith transaction and N is the number of transactions. Let θ be a minimum support threshold and μ as user specified membership function.

A quantitative database shown in Table 2 is assumed as a running example to illustrate the proposed approach. It consists of 4 transactions and 3 items denoted as (A) to (C). The minimum support threshold is initially set as θ (=30%). Fig. 1 shows the fuzzy membership function used in the example and all items in the example used the same membership function to fuzzifier the quantitative values.

Table 3
Fuzzified data from Table 1.

TID	Transformed linguistic terms
T1	(0.2/A.M + 0.8/A.H) (1.0/B.M) (0.75/C.L + 0.25/C.M)
T2	(1.0/A.H) (0.75/B.L + 0.25/B.M) (0.8/C.M + 0.2/C.H)
T3	(0.4/A.M + 0.6/A.H) (0.25/B.L + 0.75/B.M) (0.5/C.L + 0.5/C.M)
T4	(1.0/A.H) (1.0/B.M) (0.75/C.L + 0.25/C.M)
TOTAL	(0.6/A.M + 3.4/A.H) (1.0/B.L + 3.0/B.M) (2.0/C.L + 1.8/C.M + 0.2/C.H)

Table 4
Items with their frequent fuzzy regions.

Items (L_1)	Frequent regions	Fuzzifier value
A	High	3.4
B	Middle	3.0
C	Low	2.0

Definition 1. The linguistic variable R_j is an item of a quantitative database whose value is the set of fuzzy linguistic terms represented as $(R_{j1}, R_{j2}, \ldots R_{jh})$ where h is the number of fuzzy regions.

For example, in Table 2 there are three items (A), (B) and (C) and the number of linguistic terms is set as 3 that can be represented as *Low(L)*, *Middle(M)* and *High(H)*.

Definition 2. The fuzzy set (f_{ij}) is the set of fuzzy linguistic terms with their membership values transformed from quantitative value of the linguistic variable R_j as:

$$f_{ij} = (f_{ij1}/R_{j1} + f_{ij2}/R_{j2} + \ldots\ldots + f_{ijh}/R_{jh})$$

where h is the number of fuzzy regions for item j, R_{jl} is the lth fuzzy region of item j, $1 \le l \le h$ and $f_{ijl} \subseteq [0, 1]$.

For example the value "9" of (A) in the first transaction is converted into $(0.0/Low + 0.2/Middle + 0.8/High)$ by the membership functions in Fig. 1. The same process is repeated for the other items and the transformed results are shown in Table 3.

Definition 3. The count of the fuzzy region R_{jl} $(count_{jl})$ in dataset is the sum of the transformed fuzzy values that can be defined as

$$count_{jl} = \sum_{i=1}^{N} f_{ijl},$$

where N is the total number of transactions.

For example in Table 3, the fuzzy itemset (A.H) appears in transactions T1, T2, T3 and T4; the count of (A.H) can be calculated as: count (A.H) $(= 0.8 + 1.0 + 0.6 + 1.0)$ (=3.4).

Definition 4. $max\text{-}count_j$ is defined as the maximum value among all the region values for the item j and the corresponding region is assumed as the frequent region if satisfies θ.

For example, $max\text{-}count_A$ is $max(0.0, 0.6, 3.4)$ $(= 3.4)$ which corresponds to the fuzzy region *High*. Table 4 shows the frequent fuzzy regions with their values.

Definition 5. *L2Matrix* is defined as an upper triangular Boolean matrix, which shows all the combinations of frequent 1-itemsets (L_1) that form the set of frequent 2-itemsets (L_2). Every item I_i in L1 is intersected with every other succeeding item I_j and if the

Table 5
L2Matrix for the dataset.

Item/index	B.M/1	C.L/2
A.H/0	T	T
B.M/1	F	T

Table 6
Set of frequent itemsets derived from the example dataset.

Item	T1	T2	T3	T4	Total
A	0.8	1.0	0.6	1.0	3.4
B	1.0	0.25	0.75	1.0	3.0
C	0.75	0.0	0.5	0.75	2.0
AB	0.8	0.25	0.6	1.0	2.65
AC	0.75	0.0	0.5	0.75	2.0
BC	0.75	0.0	0.5	0.75	2.0
ABC	0.75	0.0	0.5	0.75	2.0

resultant itemset satisfies 'θ' the corresponding intersecting bit position in L2matrix is set as one. Meanwhile the resultant itemset with value is added into L2-itemsets.

$$L2Matrix(ij) = \begin{cases} 1, & if\ (\mathrm{I}_i \cap \mathrm{I}_j) \geq \theta \\ 0, & otherwise \end{cases}$$

where, I_i and I_j are the items in L_1 and $\mathrm{I}_i < \mathrm{I}_j$ i.e., I_i is the preceding element of I_j. Table 5 shows the L2Matrix for the example dataset.

Definition 6. Frequent k-itemsets (L_k, $k >= 2$) can be generated as follows: Let the k-itemset ($k = 2$) be AB, matrix size is $m \times m$ ($m = 3$; i.e., 3×3 Matrix with Row Keys.: A to C; Col Keys.: B to D) and assume the matrix as:

	$B(1)$	$C(2)$	$D(3)$
$A(0)$	1	1	0
$B(1)$	0	0	1
$C(2)$	0	0	1

The itemset AB is intersected only with the item C, since there is no more succeeding element has the value of '1' on the same row key 'A' and next to Col Key 'B'. Table 6 shows the set of frequent itemsets generated from the example dataset.

3.1. Fuzzy association rules

Association rules (ARs) are used to represent the dependencies between items in a database. ARs are expressions in the form A $\rightarrow$ B, where A and B are sets of items and $\mathrm{A} \cap \mathrm{B} = \emptyset$, i.e., if all the items in A exist in a transaction then all the items in B with a high probability are also in the transaction and both A and B should not have any common items. Since the classical association rules applied only on binary dataset, fuzzy association rules have been introduced for quantitative datasets.

A fuzzy association rule is of the form: $(X \in F_X) \rightarrow (Y \in F_Y)$ where $X, Y \subset I$, $X \cap Y = \emptyset$, $X = \{x_1, x_2, \ldots, x_p\}$ and $Y = \{y_1, y_2, \ldots, y_q\}$ are attributes, and $F_X = \{f_{x1}, f_{x2}, \ldots, f_{xp}\}$ and $F_Y = \{f_{y1}, f_{y2}, \ldots, f_{yq}\}$ are fuzzy sets that characterize X and Y respectively.

Definition 7. The fuzzy support count (*fsupcnt*) and fuzzy support (*fsup*) of a fuzzy association rule $(\mathrm{A}, F_\mathrm{A}) \rightarrow (\mathrm{B}, F_\mathrm{B})$ is defined as:

$$fsupcnt((\mathrm{A}, F_\mathrm{A}) \rightarrow (\mathrm{B}, F_\mathrm{B})) = \sum_{x \in T} (F_\mathrm{A} \cap_T F_\mathrm{B})(x)$$

and respective

$$fsup((\mathrm{A}, F_\mathrm{A}) \rightarrow (\mathrm{B}, F_\mathrm{B})) = \sum_{x \in T} (F_\mathrm{A} \cap_T F_\mathrm{B})(x)/|T|$$

Definition 8. The fuzzy confidence of a fuzzy association rule $(\mathrm{A}, F_\mathrm{A}) \rightarrow (\mathrm{B}, F_\mathrm{B})$ is defined as:

$$fconf((\mathrm{A}, F_\mathrm{A}) \rightarrow (\mathrm{B}, F_\mathrm{B})) = \sum_{x \in T} (F_\mathrm{A} \cap_T F_\mathrm{B})(x) / \sum_{x \in T} F_\mathrm{A}(x)$$

4. The proposed IFFP algorithm

The proposed IFFP (Improved Fuzzy Frequent Pattern Mining) algorithm for mining fuzzy frequent patterns from quantitative database is described in this section. The notation used in the proposed fuzzy frequent itemsets mining algorithm is described below.

4.1. Notations

D	–	*original quantitative database*
N	–	*total number of transactions*
M	–	*total number of items*
T_i	–	*i*th *transaction datum*, $1 \leq i \leq N$
I_j	–	*j*th *item*, $1 \leq j \leq M$
h_j	–	*number of fuzzy regions for* I_j
R_{jl}	–	*l*th *fuzzy region of* I_j, $1 \leq l \leq h_j$
V_{ij}	–	*quantitative value of* I_j *in* T_i
f_{ijl}	–	*fuzzy membership value of* V_{ij} *in region* R_{jl}
$count_{jl}$	–	*count of the fuzzy region* R_{jl} *in* D
$max\text{-}count_j$	–	*maximum count value among the fuzzy regions of* I_j
$R_{j\text{-}max}$	–	*fuzzy region of* I_j *with* $max\text{-}count_j$
θ	–	*predefined minimum support threshold*

4.2. Algorithm

The proposed IFFP algorithm integrates the fuzzy-set concepts and the improved Apriori [51–53] like approach to find fuzzy frequent itemsets from quantitative transaction data. It generates the frequent itemsets only by a single scan of the dataset. L2Matrix, an upper triangular Boolean matrix is used to reduce the vast number of candidate generation during the generation of frequent itemsets. Minimum operator is used for the fuzzy intersection operation.

The proposed algorithm consists of three phases: i) Data Transformation ii) L2Matrix formation and iii) Frequent itemsets generation. During the first phase, the algorithm transforms the quantitative values in transactions into the fuzzy set with several linguistic terms by using the given member functions. The membership value of each linguistic term in the transformed database is then summed up to determine whether the summed membership value is greater than or equal to the minimum support threshold 'θ'. Next, the transformed transactions are refined to find the satisfied linguistic terms as the frequent 1-items 'L1'. Set of frequent 1-itemsets 'L_1'is then used to construct the L2Matrix and to derive the set of frequent 2-itemsets 'L_2'. The second phase constructs the L2Matrix and generates the set of L2-itemsets.

L2Matrix is an upper triangular Boolean matrix that shows all the combinations of L1-itemsets that form the set of frequent 2-itemsets. The size of the L2Matrix is assumed to be (m-1)(m-1) where m is the number of frequent 1-itemsets 'L_1'. Let us assume the row numbers of the matrix are varied from 0 to m-2 whereas the column numbers are varied from 1 to m-1. Initially all the elements in the matrix are initialized as 'false'. Every element in L_1 is intersected with every other successive L_1 itemset to produce the L2-itemset as well as to set the corresponding intersecting position in the Boolean matrix. L2Matrix considerably reduces the number of candidate generation since the algorithm performs the

intersection operation only based on the value of the Boolean matrix during the generation of subsequent frequent itemsets.

During the third phase, subsequent frequent itemsets are generated as follows: Let us assume the array data structure is used to maintain the L_1-itemsets. For every element in L_k ($k > 1$), first item and last item are fetched out (e.g., the itemset "ABC" is fetched out as "A" as the first item and "C" as the last item) and their corresponding index positions are assigned as 'rowno' and 'colno' respectively. 'rowno' and 'colno' are the variables that indicate the starting position of the Boolean matrix from where the intersection process has to be started. The elements to be considered for the intersection operation are only those column keys which have Boolean 'True' value on the range between the columns next to the 'colno' and to the end of the column of the Boolean matrix on the row identified with 'rowno'. In this phase, the algorithm intersects every element in L_k ($k > 1$) with every satisfied key column element from L2Matrix and the resultant itemset is included into the L_{k+1} itemset if its support count exceeds the minimum support count threshold. The same process is repeated until L_{k+1} becomes null.

5. Applications of IFFP on WBCD

In this section, the algorithm IFFP is applied on WBCD to find out the core features that contribute to the presence of breast cancer. Triangular membership functions are used on WBCD to represent fuzzy sets. The membership functions for the domain values of the attributes are given in Fig. 1. As shown in Fig. 1, each attribute has three fuzzy regions; Low, Middle and High. Thus, three fuzzy membership values are produced for each attribute on the dataset.

In the first experiment, the entire dataset is analyzed to find out the behavior of features present in the breast cancer dataset. It has done by setting θ as 30% and the frequencies of all attributes are measured separately. Table 7 shows the frequencies of attributes present in WBCD dataset. It indicates that all the features except the attribute "thickness" have Low range of values, since among the 699 instances, majority of instances (65.5%) belong to benign category.

In the second experiment, both benign and malignant instances are analyzed separately for mining the contribution of each at-

Algorithm: IFFP (**I**mproved **F**uzzy **F**requent **P**attern **M**ining)

```
Purpose:  Generate fuzzy frequent itemsets
INPUT:    D - the original quantitative database
          θ - Minimum support threshold
          μ - Fuzzy membership functions
OUTPUT:   Set of fuzzy frequent itemsets
/***** Phase I - Data Transformation *****/
STEP I:   Transform the quantitative value V_ij of each item I_j in the transaction into fuzzy set
          f_ij denoted as (f_ij1/R_j1 + f_ij2/R_j2 + ..... + f_ijh/R_jh) using the user given fuzzy
          membership functions, where f_ij1 is V_ij's fuzzy membership value in R_j1, h is the number
          of fuzzy regions for I_j and R_j1 is the l^th fuzzy region of I_j, 1 ≤ l ≤ h.
STEP II:  Calculate the scalar cardinality of each fuzzy region R_j1for each items in the
          transactions as: count_j1 = Σ_{i=1}^{N} f_ij1, where N is the total number of transactions.
STEP III: Find max-count_jfrom the entire fuzzy regions for all the items I_j.
STEP IV:  Check whether the value of max-count_j exceeds the predefined minimum support count
          threshold 'N*θ'. If so, put the fuzzy region into the set of frequent fuzzy regions
          (L_1). L_1 = {R_j-max | max − count_j ≥ N * θ, 1 ≤ j ≤ M}.
STEP V:   If L_1 is not null, go to the next step. Otherwise stop process.
/***** Phase II - Construction of L2Matrix and generation of L2-itemsets *****/
STEP V:   Construct L2Matrix
          (a) Create (m-1)(m-1) L2Matrix //m is the number of frequent 1-itemsets (L_1)
          (b) Initialize FALSE to all the elements in the matrix
          (c) For each item I_i in L_1 [i ∈ 0 to m-2]
                For each successive item I_j in L_1 [j ∈ 1 to m-1]
                   Calculate I_ij = I_i ∩ I_j // Minimum operator
                          is used for intersection operation
                   if supcnt(I_ij) ≥ 'N* θ' then
                     add I_ij into L_2
                     set L2Matrix[i][j] = TRUE
STEP VI:  If L_2 are empty, stop the process; otherwise, go to the next step.
/***** Phase III - Generation of subsequent frequent itemsets *****/
STEP VII: Generate subsequent frequent itemsets
          (a) Set k=2
          (b) For each item I_i in L_k [i ∈ 0 to n-1; n is the number of frequent k-itemsets]
                set rowno = index(first item in I_1) and
                    colno = index(last item in I_1)
                for each column c = colno+1 to m-1 //m-1 is the index of last column
                     if L2Matrix[rowno][c] == TRUE then
                     calculate I_ij = I_i ∩ key-column[c]
                     if supcnt(I_ij) ≥ 'N* θ' then
                        add I_ij into L_k+1
          (c) k= k+1
          (d) Repeat step VII b) through c) until L_k becomes null.
```

Table 7
Characteristics of WBCD dataset.

Attribute	Support	Support in percentage
Thik.Middle	318.69	45.59
Mito.Low	624.75	89.38
Sizu.Low	453.75	64.91
Nor.Low	496.5	71.03
Epi.Low	384.5	55.01
Bare.Low	443.25	63.41
Bland.Low	369.00	52.79
Adh.Low	487.75	69.78
Shapu.Low	436.25	62.41

tribute in the presence of breast cancer. Table 8 and Table 9 show the frequencies of itemsets generated on malignant and benign classes respectively. To understand easily the behavior of the dataset, higher levels of frequent itemsets are listed on the table since according to the apriori [54] property of "all nonempty subsets of frequent itemsets are frequent". When analyzing these two types of frequent itemsets, the attribute 'mitoses' has the Low range of values on both types. Mitoses with Low range of values are the highest support percentage (above 70%) in malignant type of dataset and above 90% in benign type of dataset. Thus it is determined that value of mitoses is not contributed in diagnosing the breast cancer. The second attribute which has the highest support percentage in malignant class is high range of Bare Nuclei (above 60%). On the other hand, low range of the same attribute comes above 80% in benign class. This indicates, when a person has the high range bare nuclei, then there will be a more chance of breast cancer.

In order to analyze the accuracy of the association rules produced, statistical measures have been applied on the data set to find out the most frequent domain value for each attribute in the dataset. As indicated in section 2, each attribute has the domain ranges between 1 and 10. The value of 1 indicates that the person belongs to the normal condition whereas the condition goes on critical as long as the attribute domain increases. Fig. 2 shows the most frequent domain value for each attribute on both categories in WBCD dataset. From Fig. 2(a), it is seen that almost all the attributes have the domain value of either 1 or 2 which confirm the most possible ranges of domain for the person being benign category. On the other hand, Fig. 2(b) shows that almost all the attributes have the higher domain value as per the requirement of being malignant category. In contrast with this requirement, the attribute mitoses has the highest percentage of minimum most domain value. It confirms that there is no need to give more importance on the attribute 'mitoses' during the diagnosis of breast cancer.

6. Experimental evaluation

In this section, the performance of 'IFFP' is compared with the recently proposed 'CMFFP-tree' [34] and 'MFFP-tree' [47] algorithms. Three algorithms have been implemented using Java language in NetBeans IDE 6.0.1. The experiments were performed on computer with Intel(R) Core(TM)i5-3210M CPU @ 2.50 GHz processor having 6.00 GB main memory. Several experiments have been performed on different real datasets like El Nino, Adult, Iris and Heart Disease from the University of California, Irvine (UCI) Machine Learning Repository [50]. Min-Max normalization is used to normalize the dataset. Table 10 shows the properties of the dataset used.

Memory usage and runtime are considered when analyzing the algorithm with 'MFFP-tree' and 'CMFFP-tree'. The memory and runtime are measured for varying values of 'θ' on different dataset with different sizes. Fig. 3 shows the execution time of three algorithms on four datasets. As shown in Fig. 3(a)-(d), proposed 'IFFP' executes in less time. The reason for this is that the proposed algorithm avoids repeated scanning of the dataset. Once the data transformation is finished, fuzzy frequent itemsets are generated immediately through a very few intersection operation with the help of L2Matrix. On the other hand, both of the other algorithms take more time because i) each frequent itemset in the transaction should be sorted out and all the transactions are read one by one for the construction of respective tree structure, ii) maintain and process multiple linguistic term for an item, iii) each node requires to keep an array for the fuzzy values of its prefix linguistic terms and more computations are necessary to build the array-list of each item, and iv) tree structure maintains number of duplicated nodes which in turn consume more time during the construction and mining of frequent itemsets. Even though CMFFP-tree algorithm has more tree nodes, comparatively it is lower than MFFP-tree algorithm because for MFFP-tree algorithm, it sorted out the fuzzy terms in descending order of their membership values in the transaction and more nodes are thus required to build the MFFP-tree structure than CMFFP-tree.

Fig. 4 shows the memory usage of the algorithms 'MFFP-tree', 'CMFFP-tree' and 'proposed IFFP' for varying values of 'θ' on different dataset. From Fig. 4(a)-(d), we can see that 'MFFP-tree' and 'CMFFP-tree' take more memory than 'IFFP' since these algorithms require more number of tree nodes to keep the related information for producing the desired fuzzy rules. It is also observed that 'MFFP-tree' consumes more memory than 'CFFP-tree' because more nodes are required to build the MFFP-tree since all the fuzzy terms are sorted out in descending order of their membership values in the transaction. As long as the minimum support threshold increases, memory usage decreases for all the three algorithms since the large number of itemsets are maintained for lower 'θ'. When analyzing these three algorithms, proposed IFFP is superior to other two algorithms in terms of memory usage since in IFFP except frequent itemsets none others are maintained. Even though L2Matrix is used, the memory consumed by this L2Matrix is comparatively very low because i) it is a Boolean matrix ii) size of the matrix depends on the number of frequent 1-itemsets instead the size of the dataset.

7. Conclusion

In this study, a new fuzzy based approach is introduced for the mining of association rules from quantitative transactions. The efficiency of the algorithm is improved by the use of L2matrix, which reduced the vast number of candidate generation during frequent itemsets generation. The algorithm was applied on Wisconsin Breast Cancer Database to detect the core features that are associated with breast cancer. The method introduced detects breast cancer efficiently since the ineffective features are identified and removed. It is found that mitoses do not contribute in diagnosing breast cancer. Also, it is determined that bare nuclei are important features in diagnosing breast cancer. Hence, it is concluded that during the prediction of breast cancer, there is no need to give importance to the feature Mitoses and bare nuclei should be given more importance. In addition to WBCD dataset, several other real datasets like Adult, El Nino, Irish and Heart have also been used to claim the performance improvements of the 'proposed IFFP' over other recent state-of-the-art algorithms.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

Table 8
Frequent itemsets formed on malignant dataset.

Class	Rules Formed	Minimum support	Support percentage	Confidence
Malignant	If(Mito.Low)	172.75	>70%	100
	If(Mito.Low)	172.75	>60%	100
	If(Bare.High)	151.20		
	If(Mito.Low)	172.75	>50%	100
	If(Bare.High)	151.20		
	If(Epi.Middle)	136.60		
	If(Blan.Middle)	133.95		
	If(Mito.Low)	172.75	>40%	100
	If(Bare.High)	151.2		
	If(Epi.Middle)	136.60		
	If(Blan.Middle)	133.95		
	If(ShaU.Middle)	118.10		
	If(sizu.Middle || Thi.Middle)	110.75		
	If(mito.Low & ShaU.Middle || Bare.High & Mito.Low || epi.Middle & Mito.Low || blan.Middle & Epi.Middle)	97.69		

Table 9
Frequent itemsets formed on benign dataset.

Class	Rules Formed	Minimum support	Support percentage	Confidence
Benign	If(mito.Low)	412.2	>90%	100
	If(adh.Low)	419.5		
	If(norm.Low)	430.75		
	If(sizu.Low)	423.5		
	If(mito.Low & norm.Low || adh.low & mito.Low || mito.Low & sizu.Low)	415.5		
	If(adh.Low & bare.Low & mit.Low & norm.Low & shaU.Low & sizU.Low)	386.5	>80%	100
	If(adh.Low & bare.Low & blan.Low & epi.Low & mito.Low & norm.Low & shaU.Low & sizU.Low)	386.5	>70%	100
	If(adh.Low & bare.Low & blan.Low & epi.Low & mito.Low & norm.Low & shaU.Low & sizU.Low)	386.5	>60%	100
	If(adh.Low & bare.Low & blan.Low & epi.Low & mito.Low & norm.Low & shaU.Low & sizU.Low & thik.Low)	386.5	>50%	100

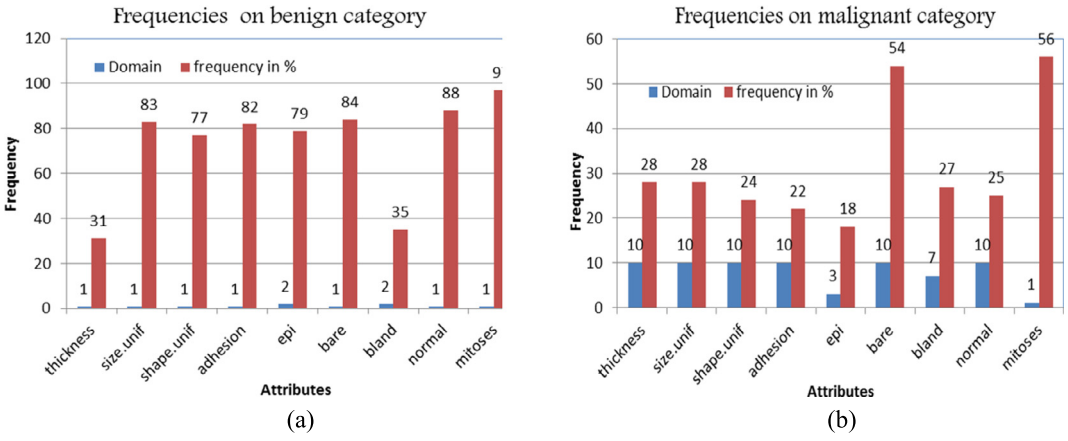


Fig. 2. Most frequent Domain values attribute wise on (a) benign category (b) malignant category.

Table 10
Datasets used in the experiments.

Datasets	Total number of transactions	Number of attributes	Number of fuzzified attributes	Types of attributes
El Nino	178080	12	36	Quantitative
Adult	48842	5	15	Quantitative
Iris	150	4	12	Quantitative
Heart disease	303	5	15	Quantitative

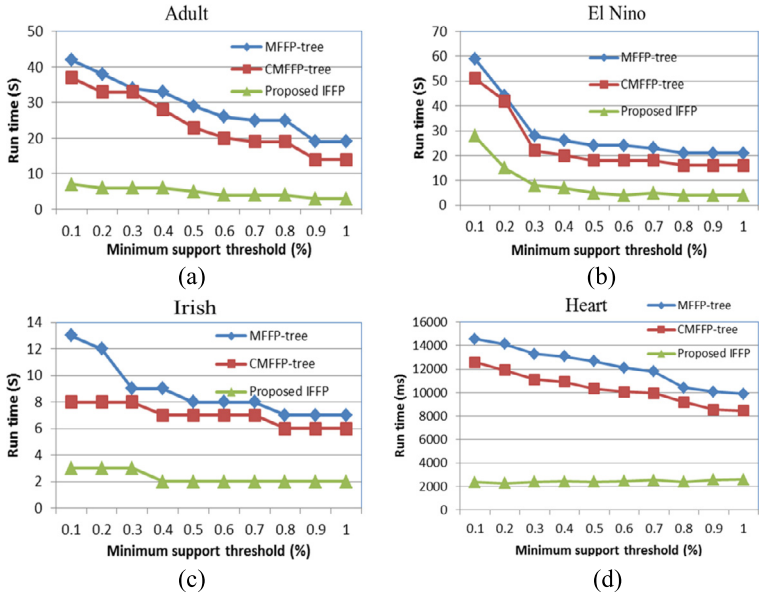


Fig. 3. The comparison of execution times among three algorithms on dataset (a) Adult (b) El Nino (c) Irish and (d) Heart disease.

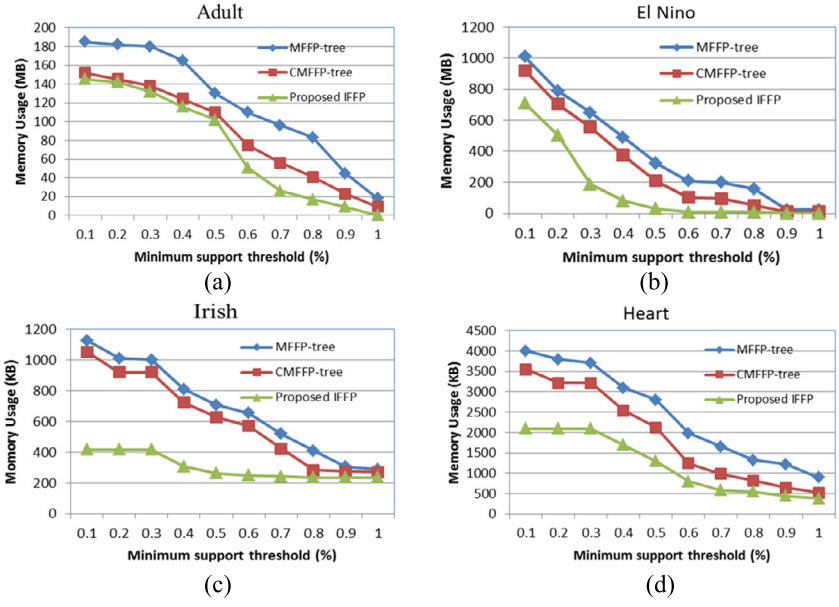


Fig. 4. The comparison of memory usage among three algorithms on dataset (a) Adult (b) El Nino (c) Irish and (d) Heart disease.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

References

[1] Jemal A, Murray T, Samuels A, Ghafoor A, Ward E, Thun MJ, et al. CA Cancer J Clin 2003;53(1):5–26.
[2] Scheidhauer K, Walter C, Seemann MD. FDG PET and other imaging modalities in the primary diagnosis of suspicious breast lesions. Eur J Nucl Med Mol Imaging 2004;31(Suppl. 1):70–9.
[3] Kenny T, Willacy H, Jackson C. About breast cancer. http://www.patient.co.uk/health/breast-cancer-leaflet, 2012. Doc. ID: 4807 (v41).
[4] Chou S-M, Lee T-S, Shao YE, Chen I-F. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst Appl 2004;27(1):133–42.
[5] Elsayad AM. Predicting the severity of breast masses with ensemble of Bayesian classifiers. J Comput Sci 2010;6(5):576–84.
[6] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. N Engl J Med 1994;331(22):1493–9. https://doi.org/10.1056/NEJM199412013312206.
[7] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M. Associations statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Syst Appl 1999;17(3):223–32.
[8] Kovalerchuck B, Triantaphyllou E, Ruiz JF, Clayton J. Fuzzy logic in computer - aided breast-cancer diagnosis: analysis of lobulation. Artif Intell Med 1997;11(1):75–85.
[9] Anderson TW. An introduction to multivariate statistical analysis. New York: Wiley; 1984.
[10] Hand DJ. Discrimination and classification. New York: Wiley; 1981.
[11] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 5th ed. Prentice-Hall; 2001.
[12] Dillon WR, Goldstein M. Multivariate analysis methods and applications. New York: Wiley. ISBN 978-0-471-08317-7, 1984.
[13] Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 2009;36(2):3465–9.
[14] Seral S, Polat K, Kodaz H, Gunes S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 2007;37(3):415–23.
[15] Marcano-Cedeno A, Quintanilla-Dominguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl 2011;38(8):9573–9.
[16] Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif Intell Med 2002;25(3):265–81.
[17] Kiyan T, Yildirim T. Breast cancer diagnosis using statistical neural networks. J Electr Electron Eng 2004;4(2):1149–53.
[18] Subashini TS, Ramalingam V, Palanivel S. Breast mass classification based on cytological patterns using RBFNN and SVM. Expert Syst Appl 2009;36(3):5284–90. Part 1.
[19] Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 1999;17(2):131–55.
[20] Polat K, Gunes S. Breast cancer diagnosis using least square support vector machine. Digit Signal Process 2007;17(4):694–701.
[21] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36(2):3240–7. Part 2.
[22] Majid A, Ali S, Iqbal M, Kausar N. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Comput Methods Programs Biomed 2014;113(3):792–808.
[23] Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Appl Intell 2009;30(1):24–36.
[24] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 2014;41(4):1476–82. Part I.
[25] Ryu YU, Chandrasekaran R, Jacob VS. Breast cancer prediction using the isotonic separation technique. Eur J Oper Res 2007;181(2):842–54.
[26] Salama GI, Abdelhalim MB, Zeid M A-e. Breast cancer diagnosis on three different datasets using multi-classifiers. Int J Comput Inf Technol 2012;1(1). 2277–0764.
[27] Lu W, Li Z, Chu J. A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. Comput Biol Med 2017;83:157–65.
[28] Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res 2018;267:687–99.
[29] Sivakumar S, Nayak SR, Vidyanandini S, Ashok Kumar J, Palai G. An empirical study of supervised learning methods for breast cancer diseases. Optik, Int J Light Electron Opt 2018;175:105–14.
[30] Peng L, Chen W, Zhou W, Li F, Yang J, Zhang J. An immune-inspired semi-supervised algorithm for breast cancer diagnosis. Comput Methods Programs Biomed 2016;134:259–65.
[31] Jafari-Marandi R, Davarzani S, Gharibdousti MS, Smith BK. An optimum ANN-based breast cancer diagnosis: bridging gaps between ANN learning and decision-making goals. Appl Soft Comput 2018;72:108–20.
[32] Alwidian J, Hammo BH, Obeid N. WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl Soft Comput 2018;62:536–49.
[33] Nahar J, Imam T, Tickle KS, Chen Y-PP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst Appl 2013;40(4):1086–93.
[34] Lin JC-W, Hong T-P, Lin T-C. A CMFFP-tree algorithm to mine complete multiple fuzzy frequent itemsets. Appl Soft Comput 2015;28:431–9.
[35] Khezri R, Hosseini R, Mazinani M. A fuzzy rule-based expert system for the prognosis of the risk of development of the breast cancer. Int J Eng, Trans A: Basics 2014;27(10):1557–64.
[36] Keles A, Yavuz U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Syst Appl 2011;38(5):5719–26.
[37] Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L. A knowledge-based system for breast cancer classification using fuzzy logic method. Telemat Inform 2017;34(4):133–44.
[38] Gilal AR, Abro A, Hassan G, Jaafar J, Rehman F. A rough-fuzzy model for early breast cancer detection. J Med Imaging Health Inform 2019;9(4):688–96.
[39] Okikiola FM, Aigbokhan EE, Mustapha AM, Onadokun IO, Akinade OA. Design and implementation of a fuzzy expert system for diagnosing breast cancer. J Adv Math Comput Sci 2019;32(1):1–14.
[40] Chan KCC, Au WH. Mining fuzzy association rules. In: The 6th international conference on information and knowledge management; 1997. p. 209–15.
[41] Hong TP, Chen JB. Finding relevant attributes and membership functions. Fuzzy Sets Syst 1999;103:389–404.
[42] Hong TP, Kuo CS, Wang SL. A fuzzy AprioriTid mining algorithm with reduced computational time. Appl Soft Comput 2004;5:1–10.
[43] Hong TP, Kuo CS, Chi SC. Trade-off between computation time and number of rules for fuzzy mining from quantitative data. Int J Uncertain Fuzziness Knowl-Based Syst 2001;9(5):587–604.
[44] Kuok CM, Fu A, Wong MH. Mining fuzzy association rules in databases. SIGMOD Rec 1998;27:41–6.
[45] Lin CW, Hong TP, Lu WH. Linguistic data mining with fuzzy FP-trees. Expert Syst Appl 2010;37:4560–7.
[46] Lin CW, Hong TP, Lu WH. Fuzzy data mining based on the compressed fuzzy FP-trees. In: IEEE international conference on fuzzy systems; 2009. p. 1068–72.
[47] Hong TP, Lin CW, Lin TC. Mining complete fuzzy frequent itemsets by tree structures. In: IEEE international conference on systems man and cybernetics (SMC); 2010. p. 563–7.
[48] Papadimitriou S, Mavroudi S. The fuzzy frequent pattern tree. In: The 9th WSEAS international conference on computers; 07/2005.
[49] Ubeyli ED. Implementing automated diagnostic systems for breast cancer detection. Expert Syst Appl 2007;33(4):1054–62.
[50] http://archive.ics.uci.edu/ml/datasets.html.
[51] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD international conference on management of data; 1993.
[52] Taihua W, Fan G. Associating IDS alerts by an improved apriori algorithm. In: Third international symposium on intelligent information technology and security informatics. IEEE; 2010. p. 478–82.
[53] Mutter S, Hall M, Frank E. Using classification to evaluate the output of confidence - based association rule mining. In: AI 2004: advances in artificial intelligence, vol. 3339. 2004. p. 538–49.
[54] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the international conference on very large data bases; 1994. p. 487–99.