Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Mining the High-Dimensional Biological Dataset Using Optimized Colossal Pattern with Dimensionality Reduction

T. Sreenivasula Reddy[1*], R. Sathya[1], Mallikarjuna Rao Nuka[2]

[1]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002, India
[2]Department of Computer Applications, Annamacharya Institute of Technology & Sciences Rajampet, YSR Kadapa, Andhra Pradesh 516115, India
Email: seenu4linux@gmail.com

Received: 9 February 2023; **Revised:** 2 March 2023; **Accepted:** 15 March 2023

Abstract: Recent years have seen a lot of attention paid to the mining of enormous itemsets from high-dimensional databases. Small and mid-sized datasets take a long time to mine with traditional algorithms since they do not include the complete and relevant information needed for decision-making. Many applications, particularly in bioinformatics, greatly benefit from the extraction of Frequent Colossal Closed Itemsets (FCCI) from a large dataset. In order to extract FCCI from a dataset, present preprocessing strategies fail to remove all extraneous characteristics and rows from the data set completely. In addition, the most current algorithms for this kind are sequential and computationally expensive. A high-dimensional dataset is pruned of all extraneous characteristics and rows using the two alternative dimensionality reduction strategies presented in this paper. Then, an optimal feature value is identified by using the Equilibrium Optimizer (EO) to identify the threshold value for reduced features. It is designed to discover mutual items and build association rules if the feature value is smaller than the frequency mining algorithm (intuitionistic fuzzy rough sets, IFRS) in conjunction with the fruit fly algorithm (FFA). If the feature value exceeds the optimal threshold, then optimized length restrictions can be used to resolve the colossal pattern (CP) mining problem (length constraints, LC). Random search is used to identify the optimal threshold values of the restrictions and extract the enormous pattern using the differential evolutionary arithmetic optimization algorithm. The experiments were carried out on twenty biological datasets that were extracted from the Machine Learning Repository of the University of California, Irvine (UCI) websites and validated the proposed models in terms of various metrics.

***Keywords*:** colossal itemsets, frequent pattern mining, intuitionistic fuzzy rough set, differential evolutionary arithmetic optimization algorithm, fruit fly, random search

MSC: 68P01

1. Introduction

It is possible to discover new rules and correlations in massive databases by doing data mining, which is the act of looking for hidden patterns and trends. As a critical phase in the knowledge detection process, it is strongly linked to

DOI: https://doi.org/10.37256/cm.5120242460

the data warehousing process, in which operational data from insufficient databases is cleaned and then merged into a single data store. Mining for often occurring patterns can be divided into three categories: sequential pattern, frequently occurring set, and graph [1]. Originally created for market basket research, which analyzes customers and discovers sets of things that are frequently purchased together, recurrent itemset mining is a vital sort of data mining. Overtone rules are used to designate the algorithm's findings that related groups of objects exist for at least a certain number of periods. Those rules of association that meet the basic support and confidence requirements are thought to be promising candidates for research. Mining common items from a large data set has a key downside [2, 3]. Computer memory will be unable to store and process the enormous number of itemsets that meet the basic support requirements. These two terms were developed as a result: CFI stands for a closed frequent itemset, whereas maximal frequent itemset (MFI) stands for constructing strategies for clustering high-dimensional data [4].

Frequent Pattern Mining has yet to be addressed, even for frequent itemset mining. This is due to one of the following reasons: There is no limit on how many times an item in a recurrent set [5] can be found in a subset of that set. There are countless designs because of the well-known property of downward closure. By introducing closed recurrent itemsets [6] and best frequent itemsets [7], this redundancy issue was greatly mitigated. A frequent pattern is maximal if and only if there is no frequent super-pattern that occurs more frequently than the frequent pattern itself. Mining results, especially for closed or maximally frequent patterns, are often explosive, despite the growing importance of many mining jobs in the real world, such as bioinformatics data processing and frequent graph pattern mining. Furthermore, larger patterns, such as longer sequences in bioinformatics, are often given higher priority in mining activities in practice because of their greater significance. Giant patterns are those that are widely supported. Most mining results contain only a small quantity of large patterns of interest, making it inefficient to wait for the method to finish "stuck" as it reaches the surface [8].

Because of the plethora of data available today, a new kind of dataset called a "high-dimensional dataset" has evolved, vastly different from transactional datasets [9]. A high-dimensional data set has fewer rows but more total features. Large, multi-dimensional datasets offer a wealth of information, but extracting that information is no easy task. In those used in bioinformatics, association rule mining (ARM) gives more weight to gigantic itemsets, which are large-sized items. They play a more significant role in the decision-making process and are crucial in many contexts [10].

Zhu et al. [11] created the idea of massive itemsets as the first to do so in a pattern-fusion-based approach. None of them have an efficient pre-processing practice because they were all implemented sequentially. Sequential pre-processing approaches used by these present algorithms lead to an exponential upsurge in the mining search area because they do not prune the dataset of all superfluous features and rows. Mining only a subset of FCCI leads to a less-than-complete set of association rules, which can have a chilling effect on your ability to make sound judgments [12, 13]. There is a high probability that FCCI extracted using the bit-wise vertical bottom up colossal (BVBUC) pattern mining approach will yield inaccurate supporting data [14-17], which in turn will result in an incorrect set of association rules that will have an impact on the quality of the decisions made. Existing algorithms' ineffective closeness checking and trimming approaches make it difficult to narrow the mining search space. Sequential methods are inefficient when trying to harvest FCCI from datasets. To address these issues, the suggested model incorporates a suitable threshold value for characteristics. Two important words, "colossal pattern mining algorithm" and "FPM," will be defined in the following sections.

2. Related work

Additionally, a more efficient PrePost algorithm [18] based on the subsume notion has been put forth, and N-lists have been employed to find frequent shut patterns [19]. The nodeset structure was then projected by Deng and Lv. To reduce the size of the database, nodesets only employ Pre (or Post), and the it should be: frequent itemset mining (FIM) procedure makes use of this structure to find patterns that occur frequently. In 2016, a new nodeset approach called DiffNodesets was suggested [20], which claims to use less memory and run more quickly. Length constraint (LC) tree layout for frequent item collections was also presented by Yun et al. [21]. It has recently been proposed that the algorithms dubbed PCP-Miner [22] be used to mine massive patterns using the -core ratio. There are no supersets for a -core pattern in the database; hence, it is considered enormous. Instead of requiring the usage of a -core ratio, BVBUC proposes an acceptable lowest support threshold and an acceptable least sum of itemsets in an itemset. In terms of

runtime, however, they are superior to BVBUC.

Colossal patterns with length limits can now be discovered using a new method called LCCP. The initial difficulty of mining large patterns with length limits was discussed [23]. To swiftly determine if a huge pattern satisfies the length limitations, two new theorems were present. To speed up the mining process, theorem-based algorithms for massive pattern mining were used to reject any patterns that did not meet the length requirements. For mining enormous patterns, min-length and max-length limitations were examined in this research. To mine huge patterns, however, which necessitate an optimal threshold value selection, it takes a long time to train.

It is only appropriate to use item postponement to mine frequent patterns if the sum of frequent 1-items is modest. This approach is inefficient if the number of frequently occurring 1-item is high. High-dimensional database problems were thus addressed by the notion known as gigantic patterns. A new approach for mining massive patterns based on core fusion was then introduced by the authors.

2.1 *Problems in existing techniques*

Nevertheless, the present pattern mining algorithms still confront the following issues:

- Singleton patterns are used in all enumeration-based mining algorithms, and it takes time to calculate these designs.
- **Huge chief memory is obligatory for actual mining:** Due to the large number of candidates that an apriori-like algorithm generates for extended patterns, it is ineffective when memory is a constraint. Interplanetary memory is needed to store applicant sets for identifying common patterns of varying size. Frequent pattern (FP) growth condenses [24] into an FP tree to avoid candidate generation.
- **Real time requests need to be high-dimensional and scalable**. For smaller datasets, a number of existing methods can be effective. As dataset size grows, existing approaches demonstrate that fit falls on core data constructions and needs adequate random access memory (RAM).
- **Manifold scans over the database.** For many current approaches, the databases are searched multiple times in order to find relevant information. The storage of intermediate findings necessitates the use of efficient data storage structures.
- This huge pattern requires a lot of help. Aside from revealing biologically important information about gene connections, these sequences are beneficial for deducing aspects of human behavior from nonhuman species. However, effective enumeration techniques outlined in many algorithms are projected [25-27].

3. Proposed scheme

In this section, we propose methods to identify an optimal feature value by using Equilibrium Optimizer (EO) to identify the threshold value for reduced features. It is designed to discover common items and build association rules if the feature value is smaller than the intuitionistic fuzzy rough sets (IFRS) in combination with the fruit fly algorithm (FFA). If the feature value exceeds the optimal threshold, then optimized length restrictions can be used to solve the LC. Random search is utilized to identify the optimal threshold values of the restrictions and extract the enormous pattern using the differential evolutionary arithmetic optimization algorithm.

3.1 *Preliminaries*

Let I be a usual of items $\{o_1, o_2, ..., o_d\}$. A subdivision of I is called a set. A business dataset D is a gathering of itemsets, $D = \{t_1, ..., t_n\}$, where $t_i \subseteq I$. For any itemset α, we signify the set of dealings that comprise α as $D_\alpha = \{i \mid \alpha \subseteq t_i \text{ and } t_i \in D\}$. Define the cardinality of an itemset α as the sum of items, it comprises, i.e., $|\alpha| = |\{o_i | o_i \in \alpha\}|$.

Definition 1. For a dataset D, an itemset α is recurrent if $\frac{(|D_\alpha|)}{(|D|)} > \sigma$, where $\frac{D_\alpha}{D}$ is named the provision of α in D, written $s(\alpha)$ and σ is the least support threshold, $0 \le \sigma \le 1$.

The set of dealings that cover a pattern is referred to as a "support set," and in this case, D_α is the support set. This study uses the term "frequent pattern" or "pattern of short" to refer to a set of often occurring things. In the case of the following two design patterns α', if $\alpha \subset \alpha'$, then α is a subpattern of α', and α' is a pattern of α.

3.2 *Robustness of colossal patterns*

Pattern fusion relies on our observations of massive patterns, and we demonstrate that in this section. Colossal patterns are known for their resilience, according to our investigation of the relationship between their support sets and those of their subpatterns. It is possible to delete a tiny number of pieces from large patterns, and the subsequent pattern would still have a large support set. There is a greater degree of robustness in larger patterns. A pattern's link to its subpattern is captured by the term "core pattern."

Definition 2 (core pattern). For a design α, an itemset $\beta \subseteq \alpha$ is said to be α τ-core pattern of α if $\frac{(|D_\alpha|)}{(|D_\beta|)} \geq \tau, 0 < \tau \leq 1$. τ is baptized the core relation.

For pattern α, let C_α be the set patterns, i.e., $C_\alpha = \{\beta \mid \beta \subseteq \alpha, \frac{(|D_\alpha|)}{(|D_\beta|)} \geq \tau\}$. In this case, it is a specific pattern. Because of the brevity, we will refer to it as a "core pattern" throughout the rest of this paper. Using the core pattern description, it is possible to formalize the robustness of a colossal pattern.

3.3 *Proposed dimensionality reduction*

In this research work, more than 15 biological datasets are used, which are explained in Section 4. Due to the large dataset, dimensionality reduction techniques are needed to improve classification accuracy. In this work, dimensionality reduction in features (DRIF) and principal component analysis (PCA) are used for the reduction process, which is explained as follows:

3.4 *DRIF*

The development of next-generation knowledge that is both brainy and automatic is required to keep up with this rapid growth. Two of the most critical issues in dealing with a large database are noise reduction and the high dimensionality of the database itself. In order to remove the superfluous features from the database, the DRIF technique was applied.

This process uses a single-scan method to find the most frequently occurring items or features. A single scan of a hash table can be used to find all instances of the different itemsets. To minimize the number of times a user must search through the database, a hash table is employed. If a data column's variance goes below the threshold, it is excluded. If the characteristics have more than three sub-ranges, select the item incidence that is equal to the threshold rate. Features with less than three sub-ranges are overlooked. For each characteristic in the UCI database, the same method is employed.

3.5 *PCA*

In order to transform potentially correlated variables, PCA employs an orthogonal transformation called independent PC. The first PC has the widest potential variance, making up as much of the data variability as feasible, shadowed by the second and third PCs. Using PCA reduces the correlation between predictor variables. It is possible to express each PC in this way:

$$PC_i = 1_{i1} X_1 + 1_{i2} X_2 \pm \ldots \pm 1_{in} X_n \tag{1}$$

where PC_i is the i-th PC and X_j is the j-th forecaster variable of X_j (i, j = 1, 2, ..., n). After the minimization of features, the threshold value is identified by using the EO. If the feature value is less than the identified threshold, then the mining

process will be carried out by Intuitionistic Fuzzy Rough Set with optimized FFA. Suppose that if the feature value is higher than the threshold value, then the colossal mining process will be carried out by optimized length constraints with a colossal pattern.

4. Threshold value identification using EO algorithm

At some point in 2020, there will be a metaheuristic algorithm known as the EO's design. Stochastic solution populations are generated to begin the optimization process by EO [28], as with other metaheuristic algorithms. A population of N particles is estimated as follows in EO:

$$X_i^d = X_{min} + rand_i^d \left(X_{max} - X_{min} \right) \tag{2}$$

where $i = 1, 2, \ldots, N$ and $d = 1, 2, ..., D$.

For each particle, *rand* represents a random vector between [0, 1], and N denotes the sum of particles. These are the dimensions' maximum and minimum values. These equilibrium possibilities were found when a specific fitness function was applied to the particles in the initial population.

4.1 *Equilibrium pool and candidates*

Equilibrium pools are used to store promising candidates in EO. As a result, the equilibrium pool contains the average of the four best-so-far particles, which will be used to update the model. A high level of exploration can be ensured by the EO's assistance with these four best-so-far possibilities. Using the average of these possibilities, it is possible to focus on regions near the optimal solution in search of a global optimum. Equilibrium pools are designed in accordance with this line of reasoning:

$$X_{(eq,pool)} = X_{eq(1)}, X_{eq(2)}, X_{eq(3)}, X_{eq(4)}, X_{eq(ave)} \tag{3}$$

$$X_{eq(ave)} = \frac{X_{eq(1)} + X_{eq(2)} + X_{eq(3)} + X_{eq(4)}}{4} \tag{4}$$

where $X_{(eq,pool)}$ is the equilibrium pool, $X_{eq(1)}$, $X_{eq(2)}$, $X_{eq(3)}$, and $X_{eq(4)}$ are the four best-so-far applicants. The $X_{eq(ave)}$ is the regular of four best-so-far applicants. In each iteration, the particles update their positions applicants.

4.2 *Exponential term*

In order for EO to uphold a suitable balance among global and local searches, the use of the term is critical.

Definition 3. The exponential term is outlined in this manner:

$$F = \exp\left(-\lambda\left(t - t_0\right)\right) \tag{5}$$

where λ is a chance vector among [0, 1], t is calculated as below:

$$t = \left(\frac{1 - iter}{Maxiter} \right)^{\left(\alpha \frac{iter}{Maxiter} \right)} \tag{6}$$

where *iter* is a continuous used to govern the local search behavior, and *Maxiter* is the all-out sum of iterations. However, t_0 is a parameter that is used to regulate exploration and exploitation.

$$t_0 = \frac{1}{\lambda}\ln\left(-\beta sign\left(r-0.5\right)\left[1-\exp\left(\lambda t\right)\right]\right)+t \tag{7}$$

This is controlled by the constant and contains a random vector in the range [0, 1]. In equation (6), the greater the value, the more powerful the exploration competence. Faramarzi et al. say that and are equivalent to 1 and 2, correspondingly. The final exponential term can be redefined as shadows [29] by substituting equation (6) for equation (5):

$$F = \beta sign\left(r-0.5\right)\left[\exp\left(\lambda t-1\right)\right] \tag{8}$$

4.3 *Generation rate*

The rate at which EO is generated is another critical consideration. The EO's ability to explore the search domain is aided by the generating rate. The formula for EO's generation rate (G) is as shown:

$$G = G_0 \exp\left(-\lambda\left(t-t_0\right)\right) = G_0 F \tag{9}$$

$$G_0 = GCP\left(X_{eq}-\lambda X\right) \tag{10}$$

$$GCP = \begin{cases} 0.5r_1 & r_2 \geq GP \\ 0 & r_2 < GP \end{cases} \tag{11}$$

Random vectors [0, 1] are represented by r_1 and r_2, respectively. The GCP is calculated as the generation control parameter in equation (10). In the end, EO's updating rule is defined as shown:

$$X = X_{eq} + \left(X-X_{eq}\right)F + \frac{G}{\lambda V}\left(1-F\right) \tag{12}$$

An equilibrium candidate chosen at random from the equilibrium pool is called X_{eq}, and V is a constant unit of measurement equal to one.

5. Condition 1: Intuitionistic fuzzy rough and optimized FFA

With the introduction of the concept of variable accuracy and the usage of presence degree as a bridge between IFS and rough set theory, International Financial Reporting Standards (IFRS) have been developed. Because of this, IFRS are able to handle a wide range of data types, including symbols, incessant values, and fuzzy values. By examining the similarities between several things, the intuitionistic fuzzy sets (IFS) relative can be discovered. Because of this, an intuitionistic fuzzy relation must be constructed that meets the two-dimensional restrictions of the IFS.

5.1 *Attribute reduction algorithm*

An intuitionistic fuzzy attribute perceptibility matrix is used to define the technique for reducing attributes.
INPUT: original choice table $S = (U, P \cup Q, V, f)$
OUTPUT: abridged intuitionistic fuzzy set R

1. One: preparing data for analysis and building an intuitive fuzzy data system.
2. The following is a standardization of data with some intuitive fuzzification:

$$\mu_{x_i r_i} = \frac{v_{x_i r_i} - \min_{i=1,2,\ldots,n}\left(v_{x_i r_i}\right)}{\max_{i=1,2,\ldots,n}(v_{x_i r_i} - \min_{i=1,2,\ldots,n}(v_{x_i r_i}} \tag{13}$$

where $\mu_{x_i r_i}$ represents the characteristic value of the object x_i in the unique fuzzy decision-making scheme for r_i, and $v(x_i r_i)$ embodies the standardized value.

The computed value $l(x_i r_i)$ lies in the range [0, 1], and furthermore, the intuitive metrics $pi_{x_i r_i}(x)$ given by the experts, the nonmembership degree $\gamma_{x_i r_i}(x)$ is intended by the formula:

$$\gamma_{x_i r_i}(x) = 1 - \mu_{x_i r_i}(x) - \pi_{x_i r_i}(x).$$

If $\mu_{x_i r_i}(x) = 1$ then $\gamma_{x_i r_i}(x) = 0$.

If $\mu_{x_i r_i}(x) = 0$, then $\gamma_{x_i r_i}(x) = 1$.

In this way, the attribute values of intuitionistic fuzzy association (IFA) can be articulated by sum pairs $\left\langle \mu_{x_i r_i} \gamma_{x_i r_i} \right\rangle$, in its place of triangular membership function to intuitionistic fuzzy (IF) of continuous qualities.

3. To compute $S(P)$ according to equation (14),

$$S(P)(x,y) = \cap\left\{r(x,y) | (x,y) \in U \times U, r(x,y) \in P\right\}. \tag{14}$$

The similarity relation of each IFA r_k is calculated according to equation (15):

$$R_k = \begin{cases} (\min\left\{\mu_{rk(x_i)}, \mu_{rk(x_j)}\right\} & r_k(x_i) \neq r_k(x_j) \\ \max\left\{\mu_{rk(x_i)}, \mu_{rk(x_j)}\right\}) & \\ (1,0) & r_k(x_i) = r_k(x_j) \end{cases} \tag{15}$$

4. To calculate $(P)(x,y)^*([x]_Q(x))$ where $[x]_Q$ is the correspondence class of decision attribute Q.
5. To calculate C_{ij} according to equation (16),

$$C_{ij} = r \in P \,|\, r(x_i, x_j) = S(P)(x_i, x_j), \omega(x_i, x_j), x_i, x_j \in U \tag{16}$$

where the disorder $\omega(x_i, x_j)$ is content with

$$\begin{aligned} &\alpha_i < \alpha_j, \beta_i < \beta_j; (\alpha_i, \beta_i) \\ &= S(P)^*\left([x_i]_Q(x_i), (\alpha_j, \beta_j)\right) = S(P)^*\left([x_i]_Q(x_i)\right) \end{aligned} \tag{17}$$

6. To calculate the weight of each IFA r_k according to equation (18),

$$\omega(r)k = \sum_{i=2}^{m}\sum_{j=1}^{i-1}\left[\frac{P}{|C_{ij}|}\right] \tag{18}$$

and calculate the *card* (r_k) value of each quality r_k.

7. In order to add the most important trait to the collection of candidate R, you must decide which one to use. If the weights are equal, the discernibility matrix is emptied of all possible combinations of attribute r_k, and a novel

discernibility matrix is rebuilt using only the attribute with the largest *card* (r_k). Repeat until all *R*-value decreases have been determined.

5.2 *Depiction of intuitionistic fuzzy rule*

The IFIS uses fuzzy rules to describe the link among symptoms and faults. A symptom can serve as a shorthand for a certain characteristic in some jurisdictions. Fuzzy matching can be used to forecast problems once symptoms have been preprocessed. Looking at the errors that have occurred, backward reasoning reveals the probable causes of the errors. Attempting to discover a problem using fuzzy correlation based on symptoms will result in a substantial amount of wasted effort. As a result, duplicate rules must be eliminated via optimization. Therefore, an intuitionistic fuzzy relation satisfying sum in [0, 1] and resemblance (x, y) is required.

5.3 *Construction of optimization perfect*

The fuzzy, rough rule base must be examined and constrained. Reducing the number of redundant rules while also increasing the number of problems that may be successfully identified within the rough fuzzy rule set are both possible outcomes of multi-objective optimization. It is for this reason that FFA is employed in this study.

5.3.1 *FFA*

Xin-She Yang found inspiration and motivation for FFA in the swarm behavior of fireflies. The firefly insect uses bioluminescence to produce short-term bangs. To a large extent, this approach is dependent on the flash's length. When employing the firefly method, three conditions must be met: It is important to remember that the appeal of a firefly is primarily gender-based. Fireflies are drawn to the flash because of its brightness. Finally, the goal function determines the brightness of a firefly. The brightness of lightning is considered an "attractive element". In turn, this has an impact on the amount of light emitted from the device. Attraction and light intensity are the two most important factors to consider while using this approach. To get things started, the FFA creates a random firefly population. An evaluation of fitness is carried out iteratively in this step, and the population is updated to the maximum number of iterations following equation (19).

$$x_i = x_i + \beta e^{\left(-\gamma d_{ij}^2\right)}\left(x_j - x_i\right) + \alpha\left(r - 0.5\right) \tag{19}$$

Firefly's current site and the firefly's tourism appeal are the first two things to consider. Finally, there's a random component in there somewhere. For each pair of fireflies i and j, we compute the Euclidean distance between them using the formula.

6. Mining algorithm

Subsequent is a description of the mining approach that combines intuitionistic fuzzy rough rule base and FFA.

6.1 *Algorithm description*

Algorithm 2. An intuitionistic fuzzy database is used as a starting point for this approach, which extracts early rules from the database.

Input: A complete rule base *RRulebase* to be optimized.

Output: Optimized rule *OPRulebase*.

1. To give an intuitive fuzzy rough rule base *RRulebase*.
2. To set related parameters and initialize initial population size *popSize* = *Sizeleft*(*RRulebaseright*) and the maximum number *MaxGeneration* of iterations is set as 1,000.
3. To initialize population, attribute reduction algorithm is used to reduce the given intuitionistic fuzzy rough

information system, delete redundant attributes and generate the initial population.

4. Calculate the completeness metrics between rules and the whole rule base.
5. Calculate the interaction metrics between rules and the whole rule base.
6. Calculate the compatibility between rules and the whole rule base.
7. To calculate the evaluation value of each individual.
8. To calculate the fitness of each individual fitness.
9. To generate the next-generation population, use the FFA.
10. To evaluate the newly generated offspring, use the fitness function of FFA.
11. Retaining top persons is the primary objective of this task. Elite populations can be replaced by current generation populations if the offspring population's total fitness is superior to that of their parents'. Other than that, nothing has changed for the affluent.
12. It has finally come to pass that the fitness of the worst individual in the elite population has been set to zero. Thus, the worst person in an elite population gets culled from the group.
13. If the maximum number of iterations has been reached, or if the threshold requirements of completion, interaction, and compatibility have been met, the cycle will be halted. To go to step (9) and generate the next population, it must return to step (9).
14. The total optimal individual is decoded and it returns to the optimal sample population.
15. Using the optimal sample population, the optimal intuitionistic fuzzy-rough rule base is mined by the method of mining rules based on fuzzy similar classes.

7. Condition 2: Optimized length constraint-based CP model

In D, the base ratio σ, the minimum LC threshold ξ, and the maximum LC Ω, the problem of extracting CP with optimized LC is defined in Definitions 4-6.

Definition 4 (min-LC with colossal pattern). Extracting the minified LC in the D database (biological database) means that all P models in D are $|P| \geq \xi$.

Definition 5 (max-LC with colossal pattern). The database delves into giant formats that control the maximum length in D and extracts all models of colossal pattern P in D with $|Q| \leq \Omega$.

According to above two definitions, the extraction problem of CP with optimized LC is defined in next Definition 6.

Definition 6 (LC with colossal pattern). Finding all CP in D, which should satisfy the min-LC and max-LC by extracting CP with optimized LC. In this research work, this optimized LC is carried out by DAOA algorithm, which is described as follows.

7.1 *The proposed differential evolutionary arithmetic optimization algorithm (DAOA)*

In this section, the working flow of projected DAOA is given as below:

Initialization phase. The optimization process (X) starts with several random solutions, when AOA [29] is used, where the X is defined in the matrix (20). In every iteration, the best solution is considered as obtained best solution.

$$X = \begin{pmatrix} x_{1,1} & \cdots & \cdots & x_{1,j} & \cdots & x_{1,n-1} & x_{1,n} \\ x_{2,1} & \cdots & \cdots & x_{2,j} & \cdots & x_{2,n-1} & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ x_{N-1,1} & \cdots & \cdots & x_{N-1,j} & \cdots & x_{N-1,n-1} & x_{N-1,n} \\ x_{N,1} & \cdots & \cdots & x_{N,j} & \cdots & x_{N,n-1} & x_{N,n} \end{pmatrix} \tag{20}$$

Phases of the proposed DAOA are primarily introduced to advance the integrative capability of the original AOA, the solution quality, and to avoid optimal local problems. This algorithm is created by introducing the differential engine (DE) working process into the traditional AOA algorithm. The DAOA method defined by the AOA was introduced to perform exploratory and exploitative searches through DE. This creates a perfect balance between research strategies

and ensures that the selected method precludes local optimism. The DAOA process includes (1) quantifying the parameters of the algorithm, (2) creating candidate solutions, (3) estimating the fitness function for optimal threshold values, (4) selecting the best solution, and (5) activating the AOA to update solutions [29] only if the specified condition is true; instead, DE is assisted in updating the solutions [29], and (6) continuing the process or providing another condition to stop.

Algorithm 3. Major steps of the proposed DAOA process:

1. Parameters of AOA a 'm' is initialized.
2. The positions for solutions are randomly initialized, where the solutions includes (Solutions: $i = 1,..., N.$).
3. Fitness values are calculated
4. **while** ($C_1 ter < M_1 ter$) **do**
5. Calculate the best solution
6. The math optimizer accelerated (MOA) and math optimizer probability (MOP) values are updated using the equations (1) and (3) presents in [28]
7. Identify the fitness function for the above solution
8. Set 1 for i^{th} solutions
9. Generate the random values (r_1, r_2, r_3) between [0, 1], when the value is less than 0.5
10. if r_1 > MOA and $r_2 > 0.5$, **then**
11. Solution's position i is updated using equation (2)'s first law in [28]
12. **else**
13. Solution's position i is updated using equation (2)'s second law in [28]
14. **end if**
15. **if** $r_3 > 0.5$, **then**
16. Solution's position i is updated using equation (4)'s first law in [28]
17. **else**
18. Solution's position i is updated using equation (4)'s second law in [28]
19. **end if**
20. if $rand < 0.5$, then
21. Solution's position i is updated using equation (5) in [28], which is called as mutation operator
22. **else**
23. Solution's position i is updated using equation (6) in [28], which is called as crossover operator
24. **end if**
25. $C_I\ ter = C_I\ ter + 1$
26. ***end while***
27. Return the best results (x).

7.2 *Defining the threshold problem between minimum and maximum LC*

The optimal threshold value is identified between min-LC and max-LC, which is defined in this section. Creation of $K + 1$ classes is carried out by defining I as LC, where i is min-LC and j is a max-LC. The threshold value (k) is required to find optimal values between the given LC by creating $K + 1$ classes, where k is defined as $\{t_k, k = 1.....K\}$.

$$\begin{aligned} C_0 &= \left\{ I_{(i,j)} | t_0 \le I_{(i,j)} \le t_1 - 1 \right\}, \\ C_1 &= \left\{ I_{(i,j)} | t_1 \le I_{(i,j)} \le t_2 - 1 \right\}, \\ &\ldots, \\ C_k &= \left\{ I_{(i,j)} | t_K \le I_{(i,j)} \le L - 1 \right\}. \end{aligned} \tag{21}$$

where, colossal pattern is defined as L and C_k is the value of k -class. Moreover, in equation (22), the multilevel threshold is defined as the maximum upgrade problem required to find the values for the optimal threshold.

$$t_1,*t_2,*,\ldots,t_k,* = \arg^{\max}_{t_1,\ldots t_k} Fit\left(t_1,\ldots,t_k\right) \tag{22}$$

7.3 *Finding optimal threshold using random search*

To find the threshold value, the values of min-LC and max-LC are categorized into various classes based on the boundary, where the classes are defined as $Class_1$, $Class_2$, ..., $Class_k$.

In this particular DAOA, random search (RS) is used to achieve optimal threshold values. The random search (RS) method is the first hyperparameter optimization (HO) optimization technique. A heuristic optimization model is what this technique belongs to. RS investigates different combinations of the optimization parameters in a manner similar to the grid search algorithm. For the sake of simplicity, let's assume the following model:

$$\max_{parm} f\left(parm\right) \tag{23}$$

The set of tuning parameters is denoted by *parm*, and the goal function to maximize is *f* (typically, the model's accuracy). The RS approach only randomly selects a sample of features from the dimensionality reduction set to examine, while the grid search method explores all options. When compared to grid search, this is different. That's why RS is preferable to grid search when there aren't many hyper-parameters to take into account. By allowing parallel optimization, this method substantially minimizes computational complexity. Hence, by utilizing RS, we may lower the suggested model's computational complexity.

The highest value is considered the best fitness function of DAOA, which is used for extracting the CP. Based on DAOA optimization's fitness function value, an optimized length constraint is processed for extracting the colossal patterns. Here, the optimal threshold value using min-LC and max-LC is identified by using optimized LC (i.e., DAOA for CP).

8. Results and discussion

The suggested approach is tested on twenty biological datasets from the ASU and UCI repositories. Table 1 lists the specifics of 20 different biological datasets that have been put to use. According to the datasets, the suggested method can be tested in a variety of ways because of the variable numbers of instances, characteristics, and classes in each dataset.

8.1 *Dataset report*

The dataset report is summarized using Table 1.

Table 1. Dataset report

Dataset name	No. of sample	No. of class	No. of feature
Prostate_GE	102	2	5,966
GLIOMA	50	4	4,434
Cervical cancer	858	2	36
Dermatology	366	6	33
Diabetic retinopathy	1,151	2	20
Vertebral(3-lass)	310	3	6
HCC Survival	165	2	49
SCADI	70	7	205

Table 1. *Continued*

Dataset name	No. of sample	No. of class	No. of feature
Leukemia	72	2	7,070
Lung_discrete	73	7	325
Lymphoma	96	9	4,026
CLL_SUB_111	111	3	11,340
Colon	62	2	2,000
Spectf-heart	267	2	44
TOX_171	171	4	5,748
nci9	60	9	9,712
Lung cancer	32	3	56
Arrhythmia	452	16	279
SPECT Heart	267	2	22
Vertebral(2-lass)	310	2	6

8.2 *Performance metrics*

R, P, $F1$ -measure and accuracy are our primary evaluation dials, and their design formulas are used as shadows in the evaluation process:

$$Precision(P) = \frac{TP}{TP + FP} \tag{24}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{25}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{26}$$

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

8.3 *Proposed evaluation*

Initially, the dimensionality reduction process is verified along with the conditions that are designed in the proposed model. By using both reduction techniques, the features are minimized and presented in Table 2.

Table 2. Experimental study of projected technique for feature reduction

Dataset name	Features	PCP-Miner	LCCP	Proposed DRIF-PCA
Cervical cancer	36	31	29	20
Dermatology	33	27	25	23
Diabetic retinopathy	20	18	16	14
Vertebral (3-Class)	6	5	4	4
Spectf-heart	44	37	32	25

Table 2. *Continued*

Dataset name	Features	PCP-Miner	LCCP	Proposed DRIF-PCA
Cervical cancer	36	31	29	20
Dermatology	33	27	25	23
Diabetic retinopathy	20	18	16	14
Vertebral (3-Class)	6	5	4	4
Spectf-heart	44	37	32	25
TOX_171	5,748	785	654	553
Leukemia	7,070	815	638	548
Lung_discrete	325	37	21	16
Lymphoma	4,026	439	376	310
CLL_SUB_111	11,340	1,658	1,010	987
Colon	2,000	259	98	84
Prostate_GE	5,966	785	524	481
GLIOMA	4,434	512	354	297
nci9	9,712	1,414	996	857
Lung cancer	56	18	14	12
Arrhythmia	279	41	33	28
SPECT Heart	22	11	8	6
HCC Survival	49	20	11	9
SCADI	205	31	22	19
Vertebral (2-Class)	6	6	5	4

In the table, it is clearly proven that the proposed DRIF-PCA effectively minimizes the number of features compared to the other two algorithms; for instance, 548 features are reduced from 7,070 whole features by using the proposed model, where LCCP reduced 638 features and PCP-Miner reduced 815 features. In these features, some important relevant features are also minimized by the existing techniques. If the dataset has fewer features, the more relevant features are also removed. For instance, a total of 27 features are removed from the dermatology using PCP-Miner, which has 34 original features. However, the LCCP removed 25 features, whereas 23 features were removed by using the proposed model. Among 20 features, the CLL_SUB_111 and nci9 datasets have overall 11,340 and 9,712 features, respectively, while the vertebral (3-class and 2-class) datasets have a very low number of overall features, i.e., only 6 features are presented. Among these six features, PCP-Miner minimized only one, which is important, and the proposed model removed two features that are not required for the process. The proposed technique in terms of recall is given in Table 3. By using the EO algorithm, a threshold value is identified to determine which conditions must be followed by the input dataset. Here, a comparison is made between these two conditions and the existing LCCP model. In order to show better analysis, the results achieved by the proposed model, which are more than 90% on various datasets, are shown in Figures 1 to 4.

Table 3. Proposed model comparative analysis in terms of recall

Dataset name	LCCP	Proposed
Cervical cancer	71.5	82.215
Dermatology	89.4	93.5
Diabetic retinopathy	82.3	90.08
Vertebral (3-Class)	80.2	90.25
Spectf-heart	77.3	87.25
TOX_171	68.1	82.25
Leukemia	62.1	71.5
Lung_discrete	67.3	74.85
Lymphoma	77.5	85.05
CLL_SUB_111	82.4	90.95
Colon	78.9	88
Prostate_GE	69.1	92.7
GLIOMA	66.5	74.2
nci9	74.5	82.15
Lung cancer	76.5	82.45
Arrhythmia	79.7	86.9
SPECT Heart	83.1	89.25
HCC Survival	74.5	86.95
SCADI	77.2	86.55
Vertebral (2-Class)	76.4	88.3

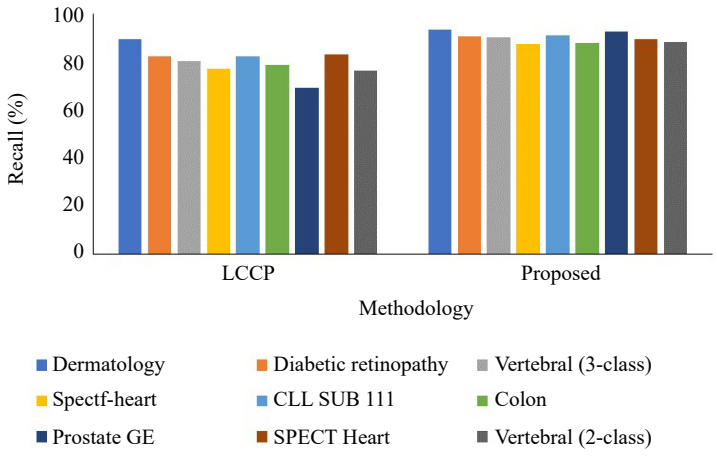

Figure 1. Graphical representation of proposed model for various dataset

Among the 20 datasets, both conditions provided low performance on three datasets such as leukemia, lung discrete, and glioma; the existing technique achieved even lower recall values (i.e., 62%-65%) than the both conditions. The existing technique, LCCP, achieved an average recall of 75.72% on 20 datasets, whereas the proposed model achieved an average recall of 85.30% on this whole dataset. The reason is that the optimal threshold value in DAOA is identified by the RS algorithm. Table 4 shows the performance validation of the proposed conditions with LCCP in terms of precision.

Table 4. Proposed model comparative analysis in terms of precision

Dataset name	LCCP	Proposed
Cervical cancer	74.2	80.2
Dermatology	88.1	93.05
Diabetic retinopathy	77.2	88.1
Vertebral (3-Class)	62.2	80.95
Spectf-heart	71.2	81.9
TOX_171	67.6	81.55
Leukemia	61.1	71.75
Lung_discrete	66.5	75.1
Lymphoma	74.4	83
CLL_SUB_111	81.8	90.95
Colon	75.9	87.7
Prostate_GE	68.6	90.15
GLIOMA	65.4	72.8
nci9	76.7	82.35
Lung cancer	76.7	82.25
Arrhythmia	77.8	87.5
SPECT Heart	82.4	90.35
HCC Survival	73.6	86.1
SCADI	76.5	89
Vertebral (2-Class)	75.2	85.8

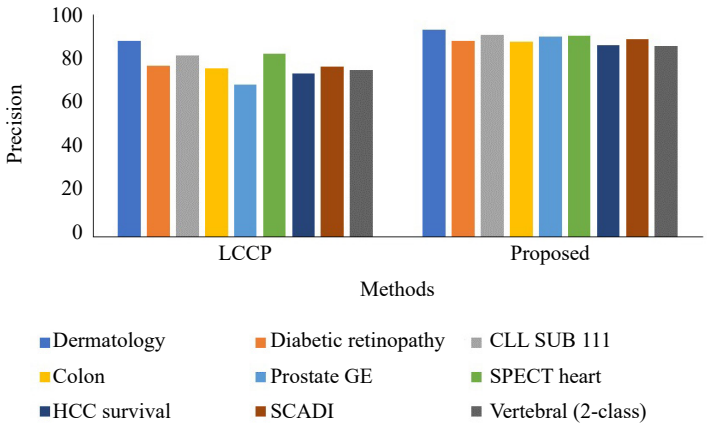

Figure 2. Graphical description of proposed model in terms of precision

In the dermatology dataset, the LCCP achieved 88.1% of *P* and the proposed model achieved 93.05%, but these models achieved only 61.1% and 71.75% on the leukemia dataset and also 66.5% and 75.1% of *P* on the lung discrete dataset. This analysis shows that features play a major role in mining patterns from a large dataset. The existing LCCP technique achieved 73.65% of average precision on all datasets, whereas the proposed model achieved 84.02% of average precision on this whole 20 datasets. Table 5 shows the performance value of the proposed conditions with the existing technique in terms of the *F*-measure.

Table 5. Proposed model comparative analysis in terms of *F*-measure

Dataset name	LCCP	Proposed
Cervical cancer	86.3	89.75
Dermatology	81.6	92.25
Diabetic retinopathy	67.2	74.15
Vertebral (3-Class)	68.0	80.65
Spectf-heart	71.8	82.7
TOX_171	65.5	78.95
Leukemia	60.1	69.25
Lung_discrete	65.6	75.55
Lymphoma	73.9	82.05
CLL_SUB_111	80.1	89.15
Colon	74.7	85.65
Prostate_GE	67.1	88.25
GLIOMA	64.1	69.35
nci9	75.3	78.4
Lung cancer	75.9	78.25
Arrhythmia	77.8	83.45
SPECT Heart	81.1	87.35
HCC Survival	72.0	84.6
SCADI	75.7	87
Vertebral (2-Class)	85.3	89.2

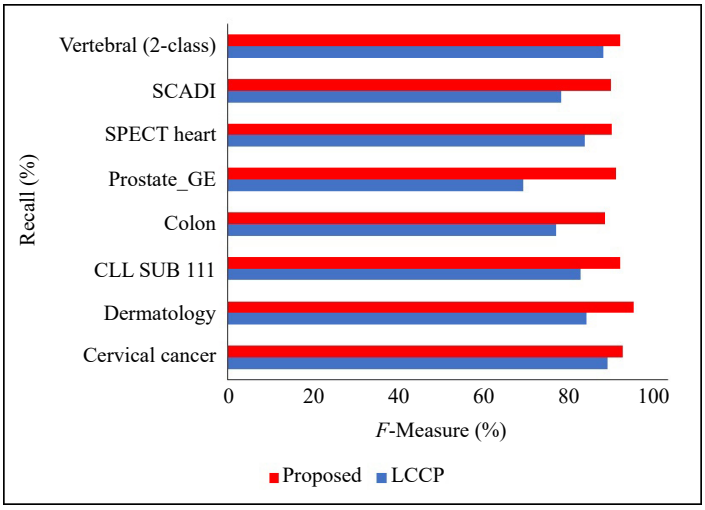

Figure 3. Graphical description of proposed conditions with existing technique

In the vertebral (2-Class), the LCCP achieved 85.3% and the proposed model achieved 89.2% of the *F*-measure. The reason for the better performance of the proposed model is that the reduced features (important features) are high in the datasets, which leads to better performance than LCCP. In addition, the LC is optimized using DAOA, and the optimal threshold value in DAOA is identified using RS. Among 20 datasets, the proposed model achieved nearly 81%

to 89% of the *F*-measure on most of the datasets and 92.25% of the *F*-measure on the dermatology dataset. The LCCP achieved an average of 73.45% of the *F*-measure, and the proposed model achieved 82.29% of the average *F*-measure on the whole 20 datasets. Table 6 provides a comparative analysis of projected conditions with LCCP in terms of accuracy on all 20 datasets.

Table 6. Proposed model comparative analysis in terms of *F*-measure

Dataset name	LCCP	Proposed
Cervical cancer	76.33	86.3
Dermatology	83.83	89.975
Diabetic retinopathy	77.57	87.935
Vertebral (3-Class)	81.45	83.615
Spectf-heart	89.13	94.455
TOX_171	76.33	87.485
Leukemia	97.92	99.82
Lung_discrete	87.86	94.4
Lymphoma	93.17	96.205
CLL_SUB_111	65.36	78.36
Colon	87.58	95.555
Prostate_GE	91.30	95.12
GLIOMA	83.90	90.92
nci9	64.17	78.68
Lung cancer	75.83	82.105
Arrhythmia	56.35	72.455
SPECT Heart	78.69	86.34
HCC Survival	68.34	76.25
SCADI	78.64	90.77
Vertebral (2-Class)	81.67	88.315

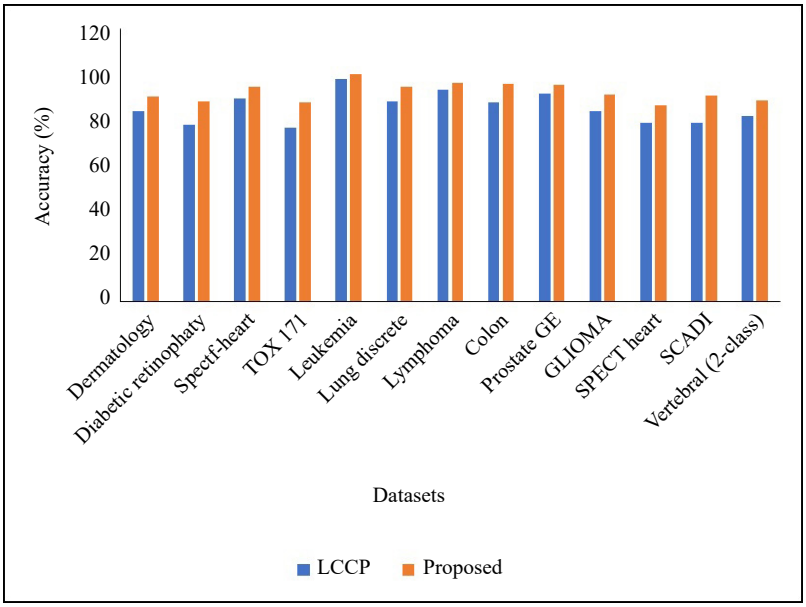

Figure 4. Experimental results of proposed conditions

The proposed model attained nearly 90% to 99% accuracy on 8 datasets among the overall dataset and nearly 80% to 89% accuracy on 8 datasets. But the existing technique, LCCP, attained nearly 91% to 97% accuracy on only 3 datasets out of 20 datasets, and this shows the better performance of the proposed model than the existing technique on all datasets. In addition, the existing LCCP is tested with only four different datasets, which is not a biological dataset, and this leads to poor performance of the LCCP model. The average accuracy of LCCP is only 79.77%, and the proposed model achieved 87.75% of the average accuracy on the overall 20 datasets.

9. Conclusion

The study includes techniques for reducing dimensions and formulating rules. To identify the most important pieces of a database with low variance, DRIF and PCA are used. According to the results of the experiments, the proposed DRIF plus PCA method exceeds all others. The best threshold value for these reduced characteristics is found using the EO method. When the feature value falls below the threshold, fuzzy ARM is used as the first condition to identify frequent items and fuzzy rules. When the feature value exceeds the ideal threshold, colossal patterns are retrieved from the twenty biological datasets using an optimal LCCP as a second requirement. An optimization process known as DAOA is used to identify the LC threshold values that are between the minimum and maximum values. The highest value is considered the final solution, which is obtained by the fitness function of DAOA with random search. A larger number of irrelevant features are effectively removed by the proposed model, where the existing techniques removed highly important features that automatically degraded its performance, as clearly shown in Tables 3 to 6. The performances of both conditions.

The proposed conditions achieved an average accuracy of 87.78% on overall datasets, whereas the existing technique achieved only 79% of the average accuracy on these datasets. In the future, this work will be improved by implementing a deep learning model for finding the optimal value to mine the patterns. In addition, a new dimensionality reduction can be included to improve its classification accuracy. Further, we implement the hyper optimization method to improve the performance.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Aqra I, Ghani NA, Maple C, Machado J, Safa NS. Incremental algorithm for association rule mining under dynamic threshold. *Applied Sciences*. 2019; 9(24): 5398. Available from: https://doi.org/10.3390/app9245398.

[2] Taşer PY, Birant KU, Birant D. Multitask-based association rule mining. *Turkish Journal of Electrical Engineering and Computer Sciences*. 2020; 28(2): 933-955. Available from: https://doi.org/10.3906/elk-1905-88.

[3] Chengyan L, Feng S, Sun G. DCE-miner: An association rule mining algorithm for multimedia based on the MapReduce framework. *Multimedia Tools and Applications*. 2020; 79: 16771-16793. Available from: https://doi.org/10.1007/s11042-019-08361-y.

[4] Reddy GT, Reddy MP, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, et al. Analysis of dimensionality reduction techniques on big data. *IEEE Access*. 2020; 8: 54776-54788. Available from: https://doi.org/10.1109/ACCESS.2020.2980942.

[5] Srinivas B, Ramesh G, Sriramoju SB. An overview of classification rule and association rule mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology.* 2018; 3(1): 1692-1697. Available from: https://ijsrcseit.com/CSEIT1831369.

[6] Wang L, Dong J-Y, Li S-L. Fuzzy inference algorithm based on quantitative association rules. *Procedia Computer Science*. 2015; 61: 388-394. Available from: https://doi.org/10.1016/j.procs.2015.09.166.

[7] Pasquier N, Bastide Y, Taouil R, Lakhal L. Discovering frequent closed itemsets for association rules. In: Beeri C, Buneman P. (eds.) *Database theory — ICDT'99*. Berlin, Heidelberg: Springer; 1999. p.398-416. Available from: https://doi.org/10.1007/3-540-49257-7_25.

[8] Kapila D, Chopra V. A survey on different fuzzy association rule mining techniques. *International Journal For Technological Research In Engineering*. 2015; 2(9): 2001-2007. Available from: https://www.ijtre.com/images/scripts/2015020957.pdf.

[9] Muyeba M, Khan MS, Coenen F. A framework for mining fuzzy association rules from composite items. In: Chawla S, Washio T, Minato S, Tsumoto S, Onoda T, Yamada S, et al. (eds.) *New frontiers in applied data mining: PAKDD 2008*. Berlin, Heidelberg: Springer; 2009. p.62-74. Available from: https://doi.org/10.1007/978-3-642-00399-8_6.

[10] Song C. Research of association rule algorithm based on data mining. In: *2016 IEEE International Conference on Big Data Analysis (ICBDA)*. China: IEEE; 2016. Available from: https://doi.org/10.1109/ICBDA.2016.7509789.

[11] Zhu F, Yan X, Han J, Yu PS, Cheng H. Mining colossal frequent patterns by core pattern fusion. In: *2007 IEEE 23rd International Conference on Data Engineering*. Turkey: IEEE; 2007. p.706-715. Available from: https://doi.org/10.1109/ICDE.2007.367916.

[12] Nguyen T-L, Vo B, Huynh B, Snasel V, Nguyen LTT. Constraint-based method for mining colossal patterns in high dimensional databases. In: Borzemski L, Świątek J, Wilimowska Z. (eds.) *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology–ISAT 2017*. Cham: Springer; 2018. p.195-204. Available from: https://doi.org/10.1007/978-3-319-67220-5_18.

[13] Nguyen T-L, Vo B, Nguyen LTT. A new method for mining colossal patterns. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Hungary: IEEE; 2016. p.003119-003124. Available from: https://doi.org/10.1109/SMC.2016.7844714.

[14] Nguyen T-L, Vo B, Snasel V. Efficient algorithms for mining colossal patterns in high dimensional databases. *Knowledge-Based Systems*. 2017; 122: 75-89. Available from: https://doi.org/10.1016/j.knosys.2017.01.034.

[15] Okubo Y, Haraguchi M. Finding top-N colossal patterns based on clique search with dynamic update of graph. In: Domenach F, Ignatov DI, Poelmans J. (eds.) *Formal concept analysis: ICFCA 2012*. Berlin, Heidelberg: Springer; 2012. p.244-259. Available from: https://doi.org/10.1007/978-3-642-29892-9_23.

[16] Sohrabi MK, Barforoush AA. Efficient colossal pattern mining in high dimensional datasets. *Knowledge-Based Systems*. 2012; 33: 41-52. Available from: https://doi.org/10.1016/j.knosys.2012.03.003.
[17] Zulkurnain NF, Haglin DJ, Keane JA. DisClose: Discovering colossal closed itemsets via a memory efficient compact row-tree. In: Washio T, Luo J. (eds). *Emerging trends in knowledge discovery and data mining: PAKDD 2012*. Berlin Heidelberg: Springer; 2013. p.141-156. Available from: https://doi.org/10.1007/978-3-642-36778-6_12.
[18] Le T, Vo B. An N-list-based algorithm for mining frequent closed patterns. *Expert Systems with Applications.* 2015; 42(19): 6648-6657. Available from: https://doi.org/10.1016/j.eswa.2015.04.048.
[19] Vo B, Le T, Coenen F, Hong TP. Mining frequent itemsets using the N-list and subsume concepts. *International Journal of Machine Learning and Cybernetics*. 2016; 7: 253-265. Available from: https://doi.org/10.1007/s13042-014-0252-2.
[20] Deng Z-H. DiffNodesets: An efficient structure for fast mining frequent itemsets. *Applied Soft Computing*. 2016; 41: 214-223. Available from: https://doi.org/10.1016/j.asoc.2016.01.010.
[21] Pyun G, Yun U, Ryu KH. Efficient frequent pattern mining based on linear prefix tree. *Knowledge-Based Systems*. 2014; 55: 125-139. Available from: https://doi.org/10.1016/j.knosys.2013.10.013.
[22] Le T, Nguyen T-L, Huynh B, Nguyen H, Hong T-P, Snasel V. Mining colossal patterns with length constraints. *Applied Intelligence*. 2021; 51: 8629-8640. Available from: https://doi.org/10.1007/s10489-021-02357-8.
[23] Bayardo RJ. Efficiently mining long patterns from databases. *ACM SIGMOD Record*. 1998; 27(2): 85-93. Available from: https://doi.org/10.1145/276304.276313.
[24] Creighton C, Hanash S. Mining gene expression databases for association rules. *Bioinformatics*. 2003; 19(1): 79-86. Available from: https://doi.org/10.1093/bioinformatics/19.1.79.
[25] Zhang Z, Teo A, Ooi BC, Tan K-L. Mining deterministic biclusters in gene expression data. In: *Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering*. Taiwan: IEEE; 2004. p.283-290. Available from: https://doi.org/10.1109/BIBE.2004.1317355.
[26] Pan F, Cong G, Tung AKH, Yang J, Zaki MJ. Carpenter: Finding closed patterns in long biological datasets. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Digital Library; 2003. p.637-642. Available from: https://doi.org/10.1145/956750.956832.
[27] Faramarzi A, Heidarinejad M, Stephens B, Mirjalili S. Equilibrium optimizer: A novel optimization algorithm. *Knowledge-Based Systems*. 2020; 191: 105190. Available from: https://doi.org/10.1016/j.knosys.2019.105190.
[28] Abualigah L, Diabat A, Sumari P, Gandomi AH. A novel evolutionary arithmetic optimization algorithm for multilevel thresholding segmentation of COVID-19 CT images. *Processes*. 2021; 9(7): 1155. Available from: https://doi.org/10.3390/pr9071155.
[29] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*. 2012; 13(2): 281-305. Available from: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.