Extraction and Classification of Best M Positive Negative Quantitative Association Rules

Ms. Sheetal Naredi, Mrs. Rushali A. Deshmukh

Abstract— Data mining an interdisciplinary research area spanning several disciplines such as expert system, database system, intelligent information systems, machine learning and statistic. In last few decades Data mining has becomes a very popular and hot area of research because of large amount of data from very large real-world database from different ecommerce websites. The data present in the datasets may be noisy and need to be processed before used. Different data cleaning techniques are used to remove this noise. This technique is used to remove the duplicates and inconsistency from the datasets. When the preprocessing on dataset is done the datasets are ready for evaluation. On the preprocessed datasets mining algorithms are applied. MOPNAR algorithm and BMPNAR algorithm are used to obtain the positive and negative values from the datasets. MOPNAR is a multi-objective algorithm for mining positive negative association rules. This algorithm might produce a large number of rules which might put more overhead on space and time resources. Hence to improve resource utilization we combine MOPNAR with Top k Rules algorithm to devise a new algorithm BMPNAR i.e. Best M positive negative Association rules algorithm. The proposed method is an extension to MOPNAR. It lets the user specify the number of rules to be generated along with the minimum confidence value. We expect the algorithm to produce Best M rules with better time and space efficiency. Once these rules are generated we need to classify them for analysis purpose. For this we make use of Firefly Algorithm. We give the comparative study of previous and new algorithm in terms of execution time and space required. The data set used is from the Keel Dataset repository.

Index Terms— Best M rules, Confidence, Datamining, Firefly Algorithm, Multiobjective Optimization, Negative association rules, Positive association rules, Rule Expansion, Support.

1 INTRODUCTION

DIFFERENT data mining concept are studied and used in several applications. Very large amount of data is used by human begins from different places every day and these data are from the different sources and fields .This data may be the documents, may be the video, may be graphical formats, may be records . Data available around us is in the different format so the datasets should be designed to take a full advantage the datasets. When and where the user will require the data should be retrieved from the database and make the better decision [1].
The method used to retrieve data from datasets is generally known as Knowledge mining or Knowledge Hub or data mining or simple KDD. Perception of “we are data rich but information poor” is the one of the important aspect that attracted a great deal of attention data industry towards field of “Data mining” to discover the useful information from the large collection. In today’s world very big number of structured and unstructured data is available but in order to use this information application we required different methods and technologies. Very huge datasets are required to generate results. To access this data efficiently and properly different techniques are used to process the data [1] [2].

Positive and negative rules are generated by Quantitative association rule mining algorithms [3]. Positive rules [4] specify the presence whereas negative rules [4] specify the absence of the property. Many more properties could be concluded based on the negative rules. While mining quantitative association rules using MOPNAR a large number of rules might be generated. These rules might be important or might not be important. Generation of large number of rules puts more overhead on the space and time requirements of the algorithm. Hence to overcome this we design an extension to MOPNAR to mine the best M positive and negative quantitative association rules. M is the number of rules to be generated specified by the user. We expect the algorithm to produce M best rules with better space and time efficiency. We combine MOPNAR and TOP k Rules Algorithm [5] to produce a new algorithm BMPNAR: Best M Positive Negative Association Rules Mining Algorithm, which mines only the Best M rules with the help of minconf value. These best M rules are then given to Firefly Algorithm [6] [7]. Firefly algorithm is a metaheuristic algorithm inspired by nature. Its working is inspired by the behavior of the Fireflies. This algorithm classifies the produced set of rules for analysis purpose. Classification uses supervised learning technique.

The remaining part of the paper is organized as follows. Section I consist of the introduction and basic challenges which motivated us for the new system. Next section II presents the related work in back ground. Section III describes the short details about the Mining Top K Association Rules. Section IV describes the Firefly Search Algorithm. The implementation detail is described in the section V. The problem statement is briefly introduced along with the methodology supported by snippet of algorithm with the details of the experimental setup and the parameter table. Section VI discusses the results and graphs.

2 RELATED WORK

Firefly algorithm [7] are nature inspired metaheuristic algorithm Firefly Algorithm. It is based on the property of firefly. The basic principle behind the algorithm is the attractiveness among the firflies due to variation in the light intensity.Due to the attractiveness property the algorithm work faster.

Mining Top-K Association Rules[5] limits the number of rules to be generated. In this paper a unique technique of mining TopK association rules has been proposed. Here K is the number of association rules to be generated. It is based on rule expansion method. It takes the help of minSup and minConf parameters to generate the top K rules.

MOEADDE [8] is multi objective evolutionary algorithm based on decomposition with differential evolution. It is an extension to MOEAD algorithm. 2 extra parameters have been used to maintain the population diversity. This algorithm is depenant on the setting of weight vector. Wrong value of weight vector may not produce optimal solution. This algorithm is not able to maintain the population diversity.

MOEA/D [9]: In this paper the author tries to optimize N single objective optimization sub problems simultaneously. It explores neighborhood relationship among the sub problems due to which the search becomes effective and efficient. The solution produced are evenly distributed with a small population. Each individual solution in the population of MOEAD is associated with the sub problem. The computation costs scales linearly with the number of decision variables.

MOPNAR [10]is an algorithm to mine positive negative quantitative association rules in a reduced set. Here the author tries to maximize three objectives Comprehensibility, Interestingness and Performance. It performs evolutionary learning of intervals of the attributes for each rule. But it might produce the large set of association rules giving less weightage to the important rules from the user perspective.

In [11]the author mines association rules based on quantitative measures. Different intervals are found by the evolutionary algorithms. This paper proposes a method to generate any number of numeric attribute in the antecedent of the rule.

R.Srikant and R. Agrawal: ”Mining quantitative association rules in large relational tables.” [3]: In this paper the author has designed a system for mining quantitative association rules from quantitative databases.

J. Alcala-Fdez, A. Fernandez, J. Luego, J. Derrac, S. Garcia, L. Sanchez et al., ”Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework.” [13]: It is a online dataset repository. The real world datasets can be downloaded. It consist of a tool to assess evolutionary algorithms for Data Mining problems.

J. Han and M. Kamber, ”Data Mining: Concepts and Techniques”, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2006. [12]: This text provides basic and detailed information about data mining. The concept of association rule is explained in details.

3 MINING TOP K ASSOCIATION RULE

Association rule mining [14] is the technique to find the rules from the datasets and it consists of discovering associations between items in transaction. This has wide applications in several domains and is used in many commercial data mining software.

The idea of mining top-k association rules used in this paper is analogous to the idea of mining top-k sequential patterns and top-k item sets . You et al. [17] consists of mining association rules from a stream instead of a transaction database whereas KORD [15] [16] only finds rules with a single item in the consequent.

The algorithm main idea is the following. Top K Rules first sets an internal minsup variable to 0. Then, the algorithm starts searching for rules. As soon as a rule is found, it is added to a list of rules L ordered by the support. The list is used to maintain the top-k rules found until now. Once k valid rules are found, the internal minsup variable is raised to the support of the rule with the lowest support in L. Raising the minsup value is used to prune the search space when searching for more rules. Thereafter, each time a valid rule is found, the rule is inserted in L, the rules in L not respecting minsup anymore are removed from L, and minsup is raised to the value of the least interesting rule in L. The algorithm continues searching for more rules until no rule are found, which means that it has found the top-k rules.

4 FIREFLY ALGORITHM

4.1 Behavior of Fireflies

Firefly algorithm first introduced by X S Yang in 2008. X S Yang has idealized some of the flashing characteristics of fireflies to develop firefly-inspired algorithms. That three idealized rules are following: [18]

- Every fireflies are unisex so that each firefly will be attracted to other fireflies regardless of their sex;
- Attractiveness is proportional to brightness of the every firefly, thus for every two flashing fireflies, the firefly with less brightness will move towards the firefly with more brightness. The attractiveness is proportional to the brightness and they both decrease as their distance increases between fireflies. If there is no brighter one than a particular firefly, it will move randomly;
- The brightness of a firefly is affected or determined by the landscape of the objective function.

4.2 Firefly Algorithm

The Firefly Algorithm is one of the newest meta-heuristics algorithms, therefore there have been written very few articles about it presented. Based on three rules stated in preceding section the pseudo-code of the basic Firefly Algorithm (FA) is illustrated in Algorithm1:

Algorithm 1: Basic Firefly Algorithm [18]

```
begin
Objective function $f(x)$;   $x = (x_1, \dots\dots, x_d)\ T$
Generate initial population of fireflies xi ($i = 1, 2..., n$)
Light intensity Ii at xi is determined by $f(x_i)$
Define light absorption coefficient g
while ($t < MaxGeneration$)
for $i = 1 : n$ all n fireflies
    for $j = 1 : i$ all n fireflies
        if ($I_i > I_j$)
        Move firefly i towards j in d-dimension
        end if
        Attractiveness varies with distance r via exp[¡g r]
        Evaluate new solutions and update light intensity
    end for j
 end for i
Rank the fireflies and find the current best
end while
Postprocess results and visualization
end
```

First each firefly generates an initial solution randomly; parameters like Light Intensity I, Initial Attractiveness β_0, and light absorption coefficient γ are defined. Then for each firefly, find the brightest firefly among them. Then the less bright firefly move towards the brightest firefly. When firefly moves or travels its light intensity decreases and its attractiveness among the other firefly will change. Then best firefly will be chosen based on an objective function for the next iteration. This condition will continue until the max iteration is reached.

5 PROPOSED METHOD

Before applying the mining to the e-commerce dataset first the dataset must be clean for processing. Preprocessing technique is used for the cleaning the dataset. Preprocessing process removes the unwanted data and repeated data from the datasets. We will be performing aggregation and duplicate remova on the unprocessed datasets. Usually association rule generated will be large in number. To limit the number of rule generated we develop system as an extension to MOPNAR to mine only the Best M rules. Depending upon the choice of parameters the current algorithm MOPNAR might generate too many rules neglecting the valuable information. Due to the limited resources it becomes expensive to handle such kind of problems. Hence our system is designed to mine Best M rules. M is the number of association rule to be generated and is specified by the user. This system uses rule expansion method and has several optimization. It limits the number of rule generated. For analysis purpose we classify the rules generated based on an attribute by applying Firefly Algorithm. Our aim is to improve the space and time efficiency

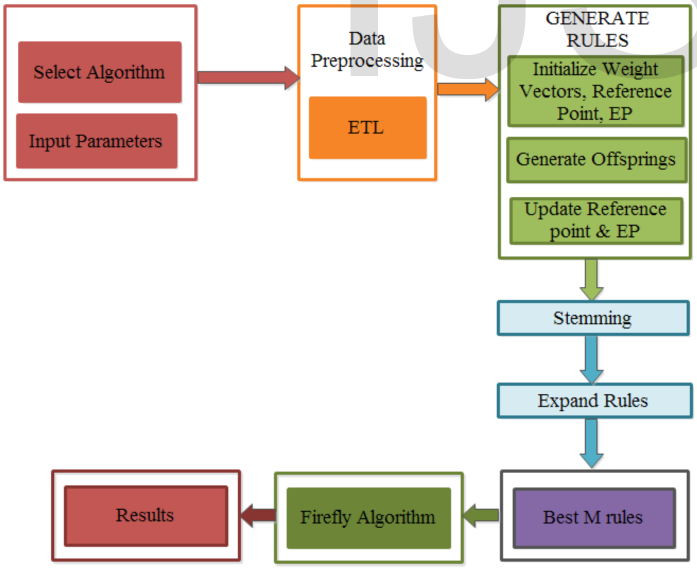


Fig.1 System Architecture

Fig.1 gives the details of the system architecture. It shows the various phases of the system. As input we provide the minimum confidence, the dataset and the additional parameter M which specifies the number of rules to be generated by BMPNAR. On the input dataset data preprocessing is done. Next is BMPNAR algorithm. It takes the initial value of minimum support and generates rules containing one item on each side i.e. on antecedent and consequent. These rules are added to a list and tested for their minconf so as to consider them for future generation. Next left and right rule expansion is applied. List of rules along with the number of rules and efficiency parameters ie. time and space required are generated. These rules will be saved in the database. The firefly algorithm takes these best M rules from the database for classification purpose. It produces a list of rules in classified form.

Hence the success definition for this system is that the system is capable of generating a limited number of classified rules as a result of which the space and time efficiency is improved.

Fig. 2 depicts the data flow architecture. The system can be made multithreaded by implementing each list with same minsup value on different threads. This makes the architecture multithreaded.

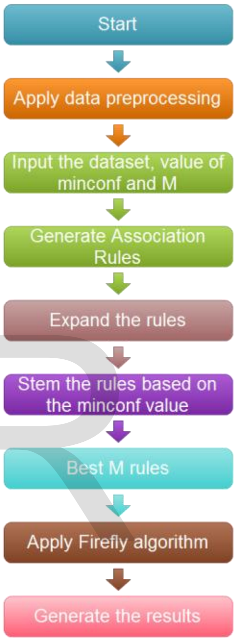

Start
Apply data preprocessing
Input the dataset, value of minconf and M
Generate Association Rules
Expand the rules
Stem the rules based on the minconf value
Best M rules
Apply Firefly algorithm
Generate the results

Fig. 2: Data Flow Architecture

5.1 ALGORITHM

The BMPNAR algorithm is used to mine the best M association rules. First the rules are generated with the minsup value. Then the rule expansion is applied to form the final list of rules. The dataset, user number of rules to be generated M and the minconf value are the inputs to the system. As the output we get the time and space required by the algorithm and number of rules generated. Once these best rules are generated they aree given to the nature inspired Firefly algorithm for classification. The final output is the Best M classified rules.

Algorithm 1 Best M Positive Negative Association Rules (BMPNAR)

```
Input: Datasets, minconf and Number of rules M.
Output: Classified Association rules.
  1: Perform data preprocessing.
  2: Initialize the weight vectors. Generate the initial population with
     N chromosomes. Initialize the reference point z and the EP.
  3: Update: For all N
```

```
        a) Generate two offsprings by crossover mutation and repairing from a solution of the population.
        b) Generate another offspring by selecting at random from the neighborhood or from population with probability ( defined by the user)
4:  Use the offsprings to update the reference point. Replace some of the solutions of the current population with worse values for the decomposition approach.
5:  These steps are repeated for each solution in population and EP is updated.
6:  Calculate the support and confidence for each rule.
7:  If the considered rule is valid it is saved in L where L is a set of current Best M rules.
8:  Each rule that is frequent is saved in R which is later considered for expansion.
9:  Generate the Best M association rules based on minconf and efficiency parameters.
10: Apply the Firefly algorithm.
        a) Generate initial population of fireflies.
        b) Define objective function.
        c) Define light absorbtion coefficient.
        d) Determine light intensity of the firefly by the objective function.
        e) For all fireflies if light intensity is greater than any other firefly move it towards that firefly by calculating the distance between them.
        f) Evaluate new solutions and update light intensity.
11: return Best M classified association rules
```

5.2 Experimental Setup

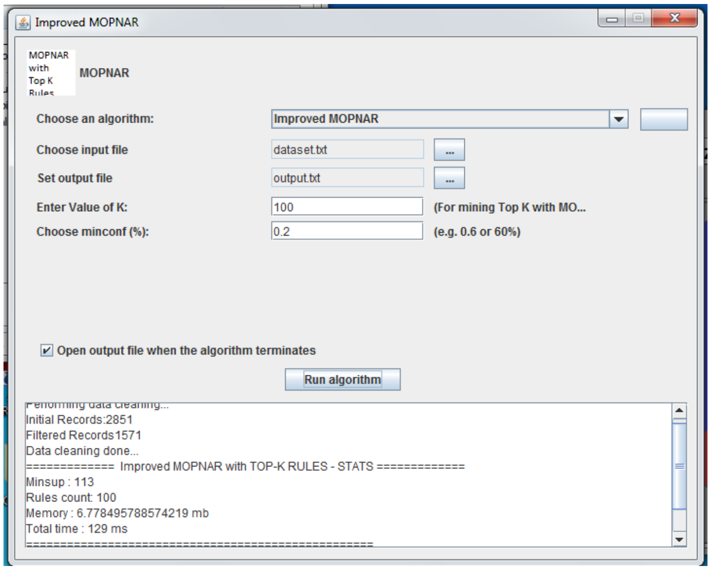
JAVA Framework (version jdk8) and Eclipse on Windows platform are used. The machine configuration on which the experiments are performed are: Intel i5 Quad core processor with 2.67 GHZ CPU, 2GB RAMS and running Microsoft Windows 7 Professional 32 bit OS.

6 Experimental Results

Figure 3 shows the GUI of the proposed work. We need to specify the type of algorithm, the dataset to be considered, the files in which output is to be stored, the minimum confidence value and the value of k. The output displays

1. The statistics for data cleaning
2. The minimum support value
3. Number of rules generated
4. Memory required
5. Time for execution.

The dataset used is of the Flipkart ecommerce website. The dataset contains the itemids of the items being purchased by the customer.
Initial testing was done using the Keel Dataset.
Table I shows the value of number of rules generated by the MOPNAR and the new algorithm BMPNAR along with their execution time and Space required. We keep the value of minconf ie. Minimum confidence value constant at 0.6. As shown in fig. 4 and fig 5 we see that as the value of M is vary; the new algorithm the BMPNAR outperforms MOPNAR in terms of execution time and the space (memory) requirement.
Fig .6 depicts the piechart. The piechart shows the distribution of the number of rules that are produced by the system after the classification algorithm is applied.



Choose an algorithm: Improved MOPNAR
Choose input file: dataset.txt
Set output file: output.txt
Enter Value of K: 100 (For mining Top K with MO...
Choose minconf (%): 0.2 (e.g. 0.6 or 60%)
☑ Open output file when the algorithm terminates

```
Performing data cleaning...
Initial Records:2851
Filtered Records1571
Data cleaning done...
============= Improved MOPNAR with TOP-K RULES - STATS =============
Minsup : 113
Rules count: 100
Memory : 6.778495788574219 mb
Total time : 129 ms
=====================================================
```

Fig 3 Snapshot

M	MOPNAR		BMPNAR	
	Time(ms)	Space(mb)	Time(ms)	Space(mb)
100	251.36	2.34	96.85	1.05
200	361.93	3.32	125.89	1.213
300	412.6	4.35	167.44	1.427
400	522.7	5.69	196.34	1.556
500	598.63	6.87	257.66	1.637
600	622.56	7.56	296.14	1.745
700	647.33	8.45	327.08	1.889
800	684.69	9.66	374.66	2.015
900	702.56	10.68	399.25	2.235
1000	746.83	12.03	413.66	2.452

Table I

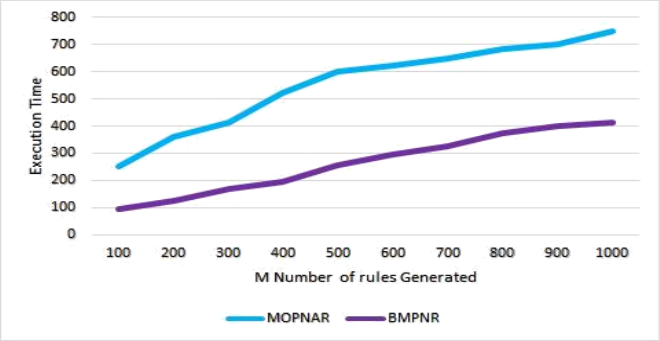


Fig. 4

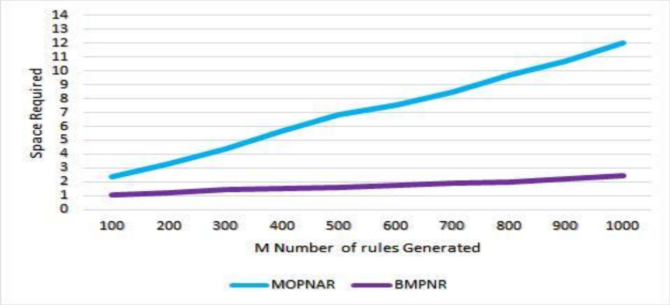


Fig.5

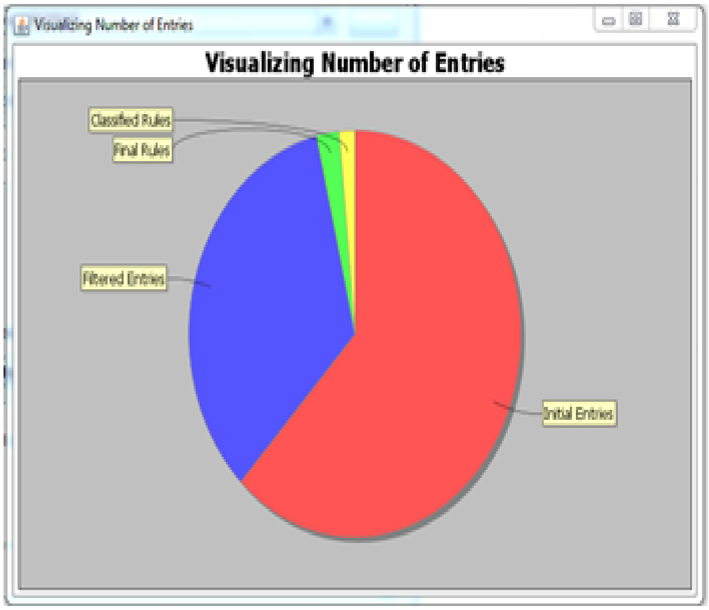


Fig 6

7 CONCLUSION

In data mining retrieving important data from the datasets is the most important task to utilize the data properly. In this paper we designed a system to mine limited M association rules which are best M rules. Instead of generating large number of rules we let the user specify the number of rules to be generated. Our aim is to improve the space and time efficiency. The rules produced are then classified by Firefly algorithm for analysis purpose.

REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi “The Survey of Data Mining Applications And Feature Scope” at International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[2] Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[3] R.Srikant and R. Agrawal, Mining quantitative association rules in large relational tables, in Proc. ACM SIGMOD, 1996, pp. 112.

[4] X.Wu, C.Zhang, and S.Zhang, Efficient mining of both positive and negative association rules, ACM Trans. Inf. Syst., vol. 22, no. 3, pp. 381405, 2004.

[5] Philippe Fournier-Viger1, Cheng-Wei Wu2 and Vincent S. Tseng2, Mining Top-K Association Rules.

[6] X.S. Yang, Firefly algorithms for multimodal optimization, SAGA 2009, Lecture Notes in Computer Science, 5792, 2009, 169-178

[7] Yang, X. S., (2010) Firefly Algorithm, Stochastic Test Functions and Design Optimisation, Int. J. Bio-Inspired Computation, Vol. 2, No. 2, pp.7884.

[8] H. Li and Q. Zhang, Multiobjective optimization problems with complicated Pareto sets, MOEA/D and NSGA-II, IEEE Trans. Evol. Computing vol. 13, no. 2, pp. 284302, Apr. 2009.

[9] Q. Zhang and H. Li, MOEA/D: A multiobjective evolutionary algorithm based on decomposition, IEEE Trans. Evol. Computing. vol. 11, no. 6, pp. 712731, Dec. 2007.

[10] Diana Martn, Alejandro Rosete, Jesus Alcala-Fdez,and Francisco Herrera, A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules,IEEE Trans. Evol. Computing, vol. 18, NO. 1, Feb 2014.

[11] J Mata, J L Alvarez, J C Riqulme : Mining Numeric Association Rules with Genetic Algorithms

[12] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2006.

[13] J. Alcala-Fdez, A. Fernandez, J. Luego, J. Derrac, S. Garcia, L. Sanchez et al., Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Computing., vol. 17, nos. 23, pp. 255287, 2011.

[14] R. Agrawal, T. Imielminski and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases,” Proc. ACM Intern. Conf. on Management of Data, ACM Press, June 1993, pp. 207-216.

[15] G I. Webb and S. Zhang, “k-Optimal-Rule-Discovery,” Data Mining and Knowledge Discovery, vol. 10, no. 1, 2005, pp. 39-79.

[16] G I. Webb, “Filtered top-k association discovery,” WIREs Data Mining and Knowledge Discovery, vol.1, 2011, pp. 183-192.

[17] Y. You, J. Zhang, Z. Yang and G Liu, “Mining Top-k Fault Tolerant Association Rules by Redundant Pattern Disambiguation in Data Streams,” Proc. 2010 Intern. Conf. Intelligent Computing and Cognitive Informatics, March 2010, IEEE Press, pp. 470-473.

[18] X. S. Yang, “Nature-Inspired Metaheuristic Algorithms”, Luniver Press, 2008.

[19] J.S. Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” *Neurocomputing—Algorithms, Architectures and Applications*, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)