COMPARATIVE ANALYSIS OF OPTIMIZATION OF ASSOCIATION RULES USING PBO AND BBO

Shelly Ahuja[1], Gurpreet Kaur[2]
Department of Computer science
Chandigarh University, Gharuan, Punjab, India

Abstract: Classification is one of the main tasks in data mining. There are various classification approaches for fetch knowledge from data. Precision rate and recall are very important in association rule and classification very important in data mining. In this work, we introduce the topics of association rules mining algorithms, classification techniques and classification based on various association rule mining. We have described a methodology for optimizing rules generated by APRIORI method. After applying optimization based on PBO, results are more optimized and efficient than BBO for classification. Parameters are evaluated for analysing the improvement in results. Main work of this paper is data set has been taken for classification. Data can be in the form of numeric values or text. Candidate generation will be performed by using APRIORI. Computation of support factor will give positive and negative rules. Positive rules will be input for PBO and negatives rules will be discarded. Rules can be changed on the basis of reproduction factor. Classification is performed on the basis of PBO and results will be compared with BBO.

Keywords: Data mining, Association Rule mining, Apriori, BBO, PBO, Classification

I. INTRODUCTION

Data Mining is a process which is used to discover pattern & relationship in data with the help of various data analysis tools to make possible prediction. Data mining is the analysis step of the Knowledge Discovery of Data process. It is the computational process of discovering patterns in huge data sets. Data Mining techniques such as classification, association and clustering are generally used to extract the hidden, previously unseen knowledge from voluminous of databases. Classification is the process which consists of predicting a particular outcome based on available input.

Steps in Classification Models

- The first step is to identify a set of subjects with a well known behaviour.
- The second step is data preparation
- The third step is training the model
- The fourth step is testing the model
- The fifth step is tweaking and refining the model

Techniques Used for Classifications in Data Mining are Decision Tree and Apriori Based Classification. Apriori is a classification algorithm for frequent item set mining and association rule learning over databases which are transactional in nature. In Apriori algorithm bottom up approach is used, where frequent subsets extends one item at a time, said to be candidate generation step, and groups of candidates are tested against the data. The algorithm itself stops when no further successful extensions are found. One of the most important rules in nature is association rule. Association rules are important features the mining process and rational selection of the effective rules is the key issues for accurate classification. The strength of the association rule is quantified by two factors i.e. support and confidence. Today, many optimization algorithms have been used in data mining. But the present available optimization algorithms don't show much effectiveness. Thus we are opting for Pollination based optimization and for showing its effectiveness its comparison has been done with Biogeography based optimization. PBO:- Optimization is a process used in the living beings. Pollination is a task of exchange of pollen from male parts of flower called anther to the female part called stigma of a flower. Flowers will introduce seeds as a conclusion of self-pollination, when pollen and pistil are producing the same plant, and from the same flower. when the pollen and pistil are from different flowers of different plants then it is said to be a cross pollination[15]. Pollinators are responsible for movement of the pollens which are further responsible for the reproduction by setting of seeds But pollinators don't have any information about the benefit of the plant. Pollination in plants is done for the energy requirement and to produce new plants. As more pollinators and their number of visits increase the pollination success rate increases.

A. POLLINATION BASED OPTIMIZATION ALGORITHM:

α=variable denoting average display at a given average nectar content. (a = optimum D = 1.2)
A = Average Investment in Nectar Content of a species (A= optimum N = 0.9. its range is 0.8 -1.4)
D = individual investment in display (0 - 1.2 typical 1.2)
N = individual investment in nectar (0.8 - 1.5 typical 1.2 at A = 0.9)
P = parameter related to pollinators learning efficiency P = m x(a+c)here m and c are
constants. (Range 0.1 – 25, typical value.2)
C =proportionality constant relating investment to reproductive cost

Step 1: Initialize PBO Parameters.

- Initial Population=Number of Plants = no. of attributes
- Number of weeks = no. of iterations

- Number of seasons = pollination value after every iteration depends on R
- Pollination weekly goal= cluster formation

Step 2: Randomly generate vectors.

- For season = 1 : number of seasons (iterations)
- For week = 1: number of weeks
- For k = 1: number of plants

Step 3: Evaluate Reproduction Vector:

$$R = \frac{(A \times D)}{(\alpha + A \times D)} + \frac{\left(\frac{a}{\alpha + A \times D}\right) \times N^P}{A^P + N^P} - C(N + D)$$

Step 4: Based on R, update number of seasons.

- Evaluate Error = Goal - R

Step 5: Based upon error update N, D, A

Step 6: Exit, if any Error acceptable.

Flow chart of Pollination based Optimization:

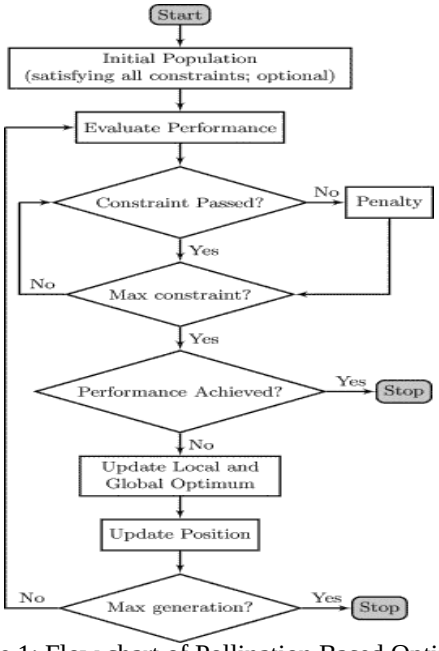

Figure 1: Flow chart of Pollination Based Optimization

B. BBO (BIOGEOGRAPHY BASED OPTIMIZATION)

Biogeography-based optimization (BBO) is an evolutionary algorithm (EA) that optimizes a function by stochastically and iteratively improving candidate solutions with regard to a given measure of quality, or fitness function. BBO belongs to the class of metaheuristics since it contain many methods, and since it does not make any assumptions about the problem and can therefore be applied to a wide class of problems. BBO is used to optimize multidimensional real-valued functions, but it cannot use the gradient of the function, it means that it does not require the function to be differentiable as required by classic optimization methods such as gradient descent and quasi-Newton methods[16]. BBO can therefore be used on discontinuous functions. Like many EAs, BBO was motivated by a natural process; in particular, BBO was motivated by biogeography, which is the study of the distribution of biological species through time and space. BBO was originally introduced by Dan Simon in 2008

Proposed Method:-

- Select the data set as input.
- By using cleaning, reduction, transformation, loading, preprocessing has been done.
- PBO(Pollination based optimization), A well known Artificial intelligence technique has been implemented for Association rule mining
- At last, Apriori algorithm is compared with optimization techniques i.e. PBO and BBO.

II. FLOW OF WORK DONE

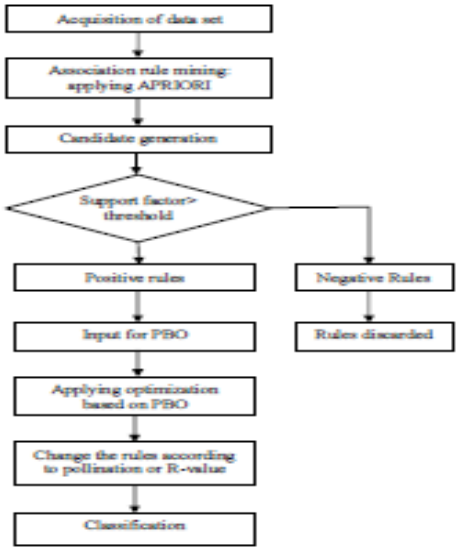

Figure 2: Flow Chart of Proposed Work

The first step is acquiring the data set of our choice. Data set can be in form of numeric or text or others. Then Association rule mining is done using apriori algorithm. Apriori algorithm basically consists of two steps. First is CANDIDATE GENERATION phase and the second is count of SUPPORT factor with THRESHOLD. If Support factor is greater than Threshold, Positive rules are generated and are accepted, otherwise Negative rules are generated and are rejected. The rules which are Positive are valid rules and rules which are declared negative by Support and Threshold are invalid .Valid i.e. Positive rules are presented as a input for Pollination based optimization. Further, rules can be changed on the basis of REPRODUCTION VECTOR (R) based on given formulae.

$$R = \frac{(A \times D)}{(\alpha + A \times D)} + \frac{\left(\frac{a}{\alpha + A \times D}\right) \times N^P}{A^P + N^P} - C(N + D)$$

On the basis of Pollination based Optimization Classification is performed and rules will be optimized and then results will be compared with BBO. After implementing PBO algorithm, we have implemented BBO algorithm and basic flow of this comparison work is given below. Rules generated by apriori algorithm are sent as a input for PBO and BBO algorithm for optimization. After getting optimized rules from both of the optimization algorithms, these results are compared by using parameters like Precision, Recall, F-Measure.

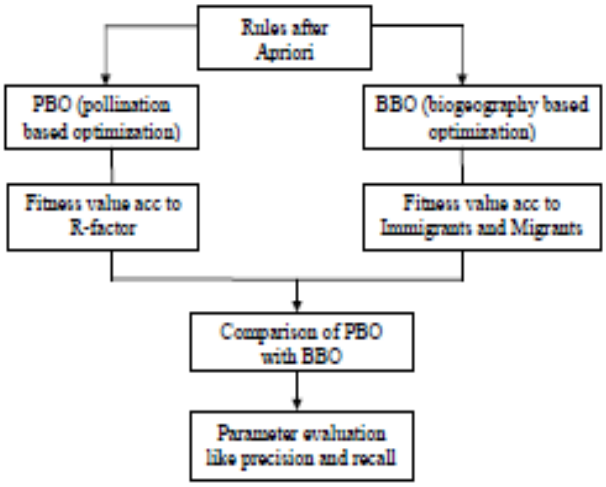

Figure 3: Flow Chart of Proposed work (part 2)

III. RESULTS

In this work, three datasets has been taken from internet. The datasets taken are: Car, Segment, Flag. Car dataset contains 2000 records. This dataset contains the attributes like number of doors, model of car, size of car etc. The dataset named Segment contains 2800 records. This dataset contains the attributes like grass, size, window, sky etc. Third dataset named Flag contains 3500 records. This dataset contains the attributes like nation, colour of flag, size of flag etc.

Name of datasets	Car[9]	Segment [10]	Flag [11]
Size of datasets	53,875	305,300	90,649
Records in a datasets	2000	2800	3500

Table 1: Table of Dataset

```
Simulation # 1/3 - PopSize = 4
Simulation # 2/3 - PopSize = 16
Simulation # 3/3 - PopSize = 64
mean min cost = 41.1711      21.4938      17.2331
mean generation count = 37.27          100          100
# samples = 45
mean improvement segment generations  = 4.2667      5.5778
mean improvement amounts = 6.0797      6.2997
# samples = 46
mean improvement segment generations  = 15.7174      16.3478
mean improvement amounts = 3.9618      5.3382      4.8117
# samples = 49
mean improvement segment generations  = 23.0612      20.4898
mean improvement amounts = 4.2717      3.4287      4.0039
```

Figure 4: Results of BBO algorithm with dataset Car showing rule generation by calculating support and confidence.

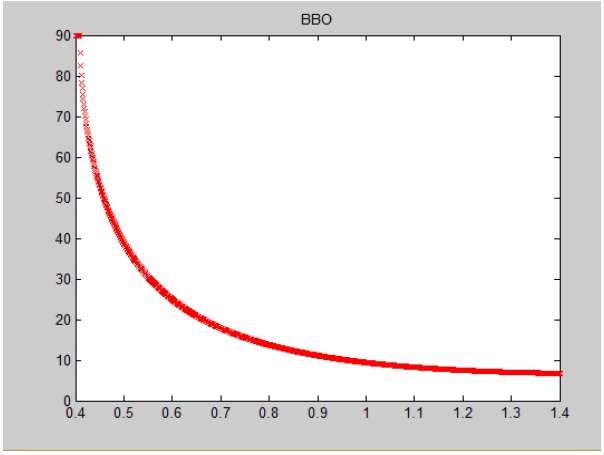

Figure 5: Graph of BBO algorithm with dataset Car showing Calculated Support and Confidence by association rules. The vertical line of graph represents Support and horizontal line represents Confidence. As the value of confidence increases, the fitness vector also increases.

```
1  4  5
1  4  6
1  4  7
2  4  5
4  5  6
1=>4  5   support  :0.66667  confidence :0.83333
4=>1  5   support  :0.66667  confidence :0.66667
5=>1  4   support  :0.66667  confidence :0.8
4  5=>1   support  :0.66667  confidence :0.8
1  5=>4   support  :0.66667  confidence :1
1  4=>5   support  :0.66667  confidence :0.83333
1=>4  6   support  :0.5  confidence :0.625
6=>1  4   support  :0.5  confidence :0.83333
4  6=>1   support  :0.5  confidence :0.83333
1  6=>4   support  :0.5  confidence :1
1  4=>6   support  :0.5  confidence :0.625
```

Figure 6: Results of PBO algorithm showing rule generation with dataset Car

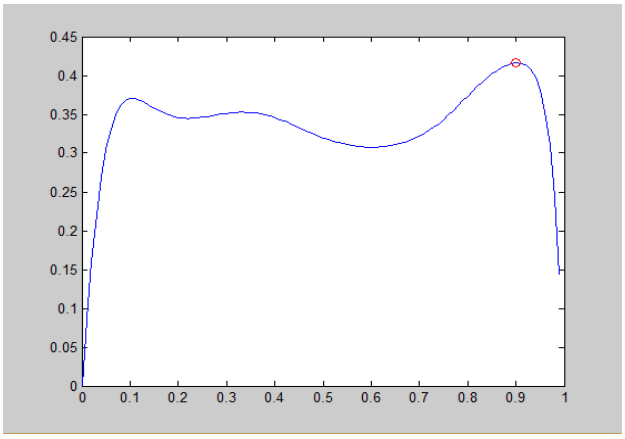

Figure 7: Graph of PBO algorithm showing the vertical line of graph represents Support and horizontal line represents Confidence. As the value of confidence increases, the fitness vector also increases.

IV. TABLE OF COMPARISON

In this table the classified data is evaluated by optimization techniques and by without optimization techniques this classification provides different values for the precision recall and f-measure by using the apriori system the apriori classification shows difference in the parameters for the different datasets. The dataset1 shows the data of the cars having all the attributes of car industry with price doors color maintenance etc. this data classify by using apriori, by using the pollination based optimization and by using the biography based optimization. The results show that pollination based optimization is best approach that can be implemented on the dataset for the classification process.

Data set name	Performance Measures	Apriori	BBO	PBO
Car	Precision	.20	.32	.41
	Recall	.29	.31	.40
	F-measure	.1634	.2910	.3190
Segment	Precision	.32	.36	.39
	Recall	.26	.34	.46
	F-measure	.4008	.6221	.7890
Flag	Precision	.49	.52	.64
	Recall	.31	.35	.40
	F-measure	.5087	.8012	.9490

Table 2: In this Comparison chart performance is measured at parameters precision, recall, F-measure. Apriori, BBO, PBO calculated at different datasets.

V. CONCLUSION

Classification is well known and an important task in data mining. In this work, the topics of traditional classification techniques association rules mining algorithms, and classification based on association rule mining and Artificial Intelligence techniques are investigated. In this work, three data sets named Car, Segment and Flag has been taken from internet and two optimization techniques have been applied on them and then evaluated for comparison. In first step Association rule mining is done using apriori algorithm. In this work rules are generated by APRIORI method and these rules are further used as a input. These input i.e. Positive rules are sent to PBO algorithm and then PBO algorithm optimizes and gives us the best rules. After applying optimization based on Pollination, results are more optimized and efficient than BBO for classification. PBO provides the best reproductive factor for the input data from which the best and fit values has been chosen for the classification process. Parameters of Precision, Recall and F-measure are evaluated for analyzing the improvement in results and these parameters shows the effectiveness of our work.

VI. FUTURE WORK

In future we will be searching the problem of creating classification systems that consider multiple classes as the antecedent and this will have great scope for future. To accomplish such task, an extensive study on text mining classification algorithms shall be made. It can be applied on huge textual data sets. By combining out this approach with other approach system will be able to handle large textual databases and others also. This technique can be used for image mining too. This classification process can be implementing for the statistical data and the huge dataset for the classification process

REFERENCES

[1] Ziyong Zhou, Yang Zhang, Integration of Association-rule and Decision Tree for High Resolution Image Classification" 2013 21st International Conference on Geoinformatics (GEOINFORMATICS), pp no. 1-4, June 2013

[2] Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I.Lakhtaria, "Efficient Classification of Data Using Decision Tree" Bonfring International Journal of Data Mining, Vol. 2, pp no. 1, March 2012.

[3] Li Fu1, Jian Cheng and Yongheng Zheng, "Object-oriented Classification of High-resolution Remotely Sensed Imagery" IPCSIT Vol. 47 (2012) © (2012) IACSIT Press, Singapore, DOI: 10.7763/IPCSIT.2012.Vol. 47.76.

[4] B VChowdar Y, Annapurna Gummadi , UNPG Raju , B Anuradha, Ravindra Changala,"Decision Tree Induction Approach for Data Classification Using Peano Count Trees" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 4, April 2012.

[5] Kishana R. Kashwan, IEEE Member, C. M. Velu, "Performance Analysis for Visual Data Mining Classification Techniques of Decision Tree, Ensemble and SOM" International Journal of Computer Applications (0975 – 8887), Vol. 57,pp no.22, November 2012.

[6] Ravindra Changala, Annapurna Gummadi, G Yedukondalu , UNPG Raju, "Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 4, April 2012.

[7] P. Rajendran, M.Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm" Journal of Computing, issue 1, ISSN 2151-9617, Vol. 2, January 2010

[8] A. Monadjemi and B. T. Thomas and M Mirmehdi, "Classification in High Resolution Images with Multiple Classifiers", Vol. 2, Issue 4, April 2012

[9] http://repository.seasr.org/Datasets/UCI/arff/car.arff .

[10] http://repository.seasr.org/Datasets/UCI/arff/segment.arff.

[11] http://repository.seasr.org/Datasets/UCI/arff/flag.arff.

[12] Usama Fayyad et.al.,"Knowledge Discovering And

Data Mining: Towards A Unifying Framework", In Proc. Of Kdd-996, July 1996.

[13] Ravindra Changala et.al. "Classification By Decision Tree Induction Algorithm To Learn Decision Trees From The Class-Labeled Training Tuples" International Journal Of Advanced Research In Computer Science And Software Engineering, Vol. 2, Issue 4, April 2012.

[14] Li Fu1 et.al.,"Object-Oriented Classification Of High-Resolution Remotely Sensed Imagery" IPCSIT,2012.

[15] Xin-She Yang et al "Flower Pollination Algorithm for Global Optimization" Springer-Verlag Berlin Heidelberg, 2012.

[16] Simon, D et.al. "Biogeography-Based Optimization" Evolutionary Computation, IEEE, Vol.12, Dec. 2008.

