HYBRIDIZATION OF ABC AND PSO FOR OPTIMAL RULE EXTRACTION FROM KNOWLEDGE DISCOVERY DATABASE

Mrs.K.JAYAVANI,
Research Scholar,
Manonmaniam Sundaranar University,
Tirunelveli.
Vanigopinath@gmail.com

Dr.G.M.KADHAR NAWAZ,
Director,
Dept.Of.MCA
Sona College Of Engineering and Tech,
Salem.

***Abstract*---Knowledge discovery in database (KDD) has provided a large interest in statistics, machine learning, and artificial intelligence (AI). It is a challenging task for mining the comprehensive and informative knowledge in such complex data by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining. In order to extract knowledge from different datasets, we will propose a hybrid mining technique. The knowledge extraction can be done by association rule mining with the combination of Artificial Bee Colony optimization algorithm (ABC) and Particle swarm optimization (PSO). The main aim of this hybridization is to extract the optimal rules from the association rules for further classification. The accuracy will be checked in terms of optimal rule obtained from the hybridization. The best position for moving the particle will be updated by using ABC algorithm.**

***Keywords:* KDD, PSO, ABC, AI.**

I. INTRODUCTION:

Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition/AI, visualization, and high-performance and parallel computing. In this paper, which is intended to be strictly a companion reference to the invited talk, and not a presentation of new technical contributions, we outline the basic notions in this area and show how data mining techniques can play an important role in the analysis of scientific data sets .

One of the important tasks in data mining is classification. In classification, there is a target variable which is partitioned into predefined groups or classes. The classification system takes labeled data instances and generates a model that determines the target variable of new data instances . The discovered knowledge is usually represented in the form of if–then prediction rules, which have the advantage of being a high level, symbolic knowledge representation, contributing to the comprehensibility of the discovered knowledge . The discovered rules can be evaluated according to several criteria, such as the degree of confidence in the prediction, classification accuracy rate on unknown-class instances, and interpretability. Accuracy and interpretability are two important criteria in data mining .

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both . Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified . Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Classification plays an important role in data mining and the need for building classifiers across multiple databases is driven by applications from various domains. Classification is about grouping data items into classes (categories) according to their properties (attribute values). Classification is also called supervised classification, as opposed to the unsupervised classification (clustering). ‘‘Supervised” classification needs a training dataset to train (or configure) the classification model, a validation dataset to validate (or optimize) the configuration, and a test dataset to evaluate the performance of the trained model. Classification methods include, for example, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbor methods and case-based reasoning (CBR). Data mining adopted its techniques from many research areas including statistics, machine learning, association rules, neural networks, and so on.

A. Association rules. Association rule generators are a powerful data mining technique used to search through an entire data set for rules revealing the nature and frequency of relationships or associations between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behavior.

B. Artificial Neural Networks: These are recognized in the automatic learning framework as universal approximations, with massively parallel computing character and good generalization capabilities, but also as black boxes due to the difficulty to obtain insight into the relationship learned.

C. Statistical Techniques. These include linear regression, discriminate analysis, or statistical summarization.

D. Machine learning (ML) is the center of the data mining concept, due to its capability to gain physical insight into a problem, and participates directly in data selection and model search steps. To address problems like classification (crisp and fuzzy decision trees), regression (regression trees), time-dependent prediction (temporal trees), and the ML field is basically concerned with the automatic design if then rules similar to those used by human experts. Decision tree induction: the best known ML framework was found to be able to handle large-scale problems due to its computational efficiency, to provide interpretable results, and, in particular, able to identify the most representative attributes for a given task .

Successful Knowledge Discovery and Data Mining applications play an important role in data that have clearly grown to surpass raw human processing abilities. The challenges facing advances in this field are formidable. Some of these challenges include as follows: Develop new mining algorithms for classification, clustering, dependency analysis, and change and deviation detection that scale to large databases .

II. KNOWLEDGE EXTRACTION USING ASSOCIATION RULE MINING:

Apriori is a classic algorithm for learning association rules. Apriori algorithm is designed to operate on databases containing transactions. The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. It was developed by Developed by Agrawal and Srikant 1994. It is an innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. In association rule mining the input given is the database. There are two main steps in association rule mining. First the frequent item sets are generated using the minimum support value assigned. Second, the association rules are generated using the frequency item sets generated and the minimum confidence value assigned.

A database is given as input to the association rule mining process. From the input database the number of transactions is calculated. For extracting association rules minimum values of support and confidence should be assigned. The item sets are extracted from the database. Each item is the member of set of candidate. The support values are calculated for the item sets separately. The support value is calculated using the formula

$$Support\,(A \rightarrow B) = P\left(A \cup B\right) \quad \textbf{(1)}$$

Where A and B = Frequent item sets.

The support values calculated for the separate item sets are compared with the minimum support value. The item sets with support value less than the minimum support value is eliminated. The remaining item sets are selected. Then the selected item sets were combined with the same item sets. Again the support value is calculated for the item sets and they are eliminated based on the support value. By the elimination and the pruning step the item set which is for generating association rules are found out. The confidence value can be found out by using the formula

$$Confidence\,(A \rightarrow B\,) = \frac{P\left(A \cup B\right)}{P\left(A\right)} \qquad \textbf{(2)}$$

Where A and B = Frequent item sets.

A. Pseudo code for Association rule mining:

```
$C_k$: Candidate item set of size k
$L_k$ : frequent item set of size k
$L_k$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do
            increment the count of all candidates in $C_{k+1}$
            that are contained in $t$
    $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
    end
return C$_k$ $L_k$;
```

The frequent item set generated finally and the minimum confidence value is used to generate association rules. All the item sets generated were taken. The item sets with the items in the frequent item set is identified. The confidence values for the selected item sets were calculated using formula (2). The item sets with confidence value greater than the minimum confidence value assigned are selected. The remaining item sets are rejected. The selected item sets are the association rules.

The association rules generated is given as input to the hybrid ABC and PSO algorithm for optimization.

III. OPTIMIZATION USING ABC AND PSO ALGORITHM:

ABC is an algorithm which is explained by Dervis Karaboga in 2005, inspired by the smart behavior of honey bees. The colony of artificial bees has three set of bees in ABC algorithm; they are employed bees, onlookers and scouts. A bee which is waiting on the dance area for making a choice to pick a association rule is called onlooker and a bee which goes to the association rules that is selected by the onlooker is called employed bee. The other type of bee is scout bee that carries out unsystematic search for discovering novel sources. The position of the association rules denotes a realistic solution to the optimization issue and the value of a association rules related to the quality (fitness) of the associated solution, estimated by,

$$FIT_i = \frac{1}{1 + r_i} \qquad \textbf{(3)}$$

Where
i = number of association rules
r = association rules

The collective intelligence of honey bee swarms consists of three components. They are Employed bees, Onlooker bees and Scout bees. There are two major behaviors.

1) Association rules:
To select the association rules forager bee evaluates various properties. For simplicity one quality can be considered.

2) Employed bees:
The employed bee is employed on a particular association rule. It shares the information about the particular association rules with other bees in the hive. The information which is carried by the bee includes direction, Profitability and the distance.

3) Unemployed bees:
The unemployed bees include both onlooker bees and the scout bees. The onlooker bee searches the association rules with the information given by the employed bees. The scout bee searches the association rules randomly from the environment.

The main steps of ABC algorithm are:

- Initialize Association rules
- repeat
- Place the employed bees on the association rules
- Place the onlooker bees on the association rules depending on their nectar amounts
- Send the scouts to the search area for discovering new association rules
- Memorize the best association rule found so far until requirements are met

In the ABC algorithm each cycle consists of three steps. Initial step involves sending the employed bee to find out the Association rules to evaluate their values then the Association rules were selected by the onlooker based on the information given by the employed bee then the scout bees was send to find the new Association rules. At the initialization stage a number of Association rules were determined by the bees and their values were calculated. At the first step of the cycle the employed bees come in to the hive and share the information about the Association rules and their value information. The information is shared by the employed bees with the bees waiting in the dance area. The onlooker bees take the information about the Association rules. Then the employed bees travels to their respective Association rules which they have already visited and finds the neighboring Association rules in comparison through visual information.

At the second step of the cycle the onlooker bee selects the Association rules depending on the information given by the employed bees. If the optimization increases the probability of the association rules chosen also increases. When the onlooker bee arrives in the area as per the information given by the employed bee it chooses the neighboring association rules by comparing the values by visual information same as in employed bees. The new association rules were found by the bees on comparison of values based on the visual information. At the third step when the association rules was taken by the bees' new association rules was found out by the bees. A scout bee randomly selects the new association rules and replaces the old association rules with the new one. The bees which has the fitness values as good enough is the result of this fitness. The detailed explanation of the ABC algorithm is as follows:

- Initialize the association rules of the solutions $s_{i,j}$.
- Calculate the population.
- Set $cycle = 1$; the cycle denotes an iterative value.
- Create a solution $u_{i,j}$ in the neighborhood of $s_{i,j}$ using the following formula:

$$u_{i,j} = s_{i,j} + \Phi_{i,j}\left(s_{i,j} - s_{k,j}\right) \qquad \textbf{(4)}$$

Where,

k → Solution of i
Φ → Random number of range [-1,1].

- Apply the greedy selection process amid $u_{i,j}$ and $s_{i,j}$ based on the fitness.
- Calculate the probability values P_i for the solutions $s_{i,j}$ using their fitness values based on the following formula:

$$P_i = \frac{FIT_i}{\sum_{i=1}^{SN} FIT_i} \qquad \textbf{(5)}$$

- In order to estimate the fitness values of the solution we have used the following formula:

$$FIT_i = \begin{cases} \frac{1}{1+r_i}, if\ r_i \geq 0 \\ 1 + abs(r_i), if\ r_i \leq 0 \end{cases} \qquad \textbf{(6)}$$

- Normalize the P_i values into [0, 1].
- Create the novel solutions $u_{i,j}$ for the onlookers from the solutions s_i depending on P_i and calculate them.
- Apply the greedy selection procedure for the onlookers amid s_i and u_i based on fitness.
- Determine the abandoned solution (source), if exist, replace it with a novel unsystematically produced solution s_i for the scout using the following equation:

$$s_{i,j} = \min{}_j + rand(0,1) \times \left(\max{}_j - \min{}_j\right) \qquad \textbf{(7)}$$

- Memorize the optimum association rules position (solution) attained so far.
- Cycle=cycle+1
- Until, cycle=maximum cycle number.

A. Pseudo-code for ABC algorithm:

Require: Max_Cycles, Colony Size and Limit

- Initialize the Association rules
- Evaluate the Association rules
- Cycle=1
- **while** $cycle \leq Max\_cycles$ **do**
- Produce new solutions using employed bees
- Evaluate the new solutions and apply greedy selection process
- Calculate the probability values using fitness values
- Produce new solutions using onlooker bees
- Evaluate the new solutions and apply greedy selection process
- Produce new solutions for onlooker bees
- Apply Greedy selection process for onlooker bees
- Determine abandoned solutions and generate new solutions randomly using PSO in scout bee section
- Memorize the best solution found so far
- Cycle = Cycle + 1
- **end while**
- **return** best solution

ABC algorithm has several dimensional search spaces in which there are Employed bees and Onlookers bees. Both bees were categorized by their experience in identifying the association rules. The initial population is opted from the employed bee phase. The rules are possessed by the employed bee. The solution of the employed bee is altered in the onlooker bee stage based on the following formula:

$$u_{i,j} = s_{i,j} + \Phi_{i,j}\left(s_{i,j} - s_{k,j}\right) \quad \textbf{(8)}$$

Where,

$s_{i,j}$ → Solution obtained from the employed bee phase

$\Phi_{i,j}$ → Randomly produced number of range [-1, 1]

k, j → Random indexes in the solution matrix of employed bee

A solution is created based on the formula and the solution is applied in the fitness function to obtain the fitness value. This process would last until the entire employed bee gets processed. The scout bee phase is the eventual stage of the ABC algorithm. This stage is implemented with PSO algorithm in order to find the optimal rule. The scout bee initiates the process by choosing the solution from the onlooker bee phase which poses the lowest fitness value. The onlooker bee phase generates diverse solution based on different $u_{i,j}$ values. The solution with least fitness value is selected.In the scout bee section PSO optimization algorithm is included.

B. PSO in scout bee phase:

The Particle swarm optimization (PSO) is a population based stochastic optimization technique. It was developed by Dr. Eberhart and Dr. Kennedy in 1995. This technique was inspired by social behavior of bird flocking or fish schooling. PSO model constructed based on three ideas. They are

- Evaluation
- Comparison
- Imitation

```
For each particle
    Initialize particle
END
Do
    For each particle
        Calculate fitness value
        If the fitness value is better than the best fitness value (pbest) in history
            Set current value as the new pbest
    End
    Choose the particle with the best fitness value of all the particles as the gbest
    For each particle
        Calculate particle velocity according Eqn. 9
        Update particle position according Eqn. 10
    End
While maximum iterations or minimum error criteria is not attained
```

In this process the potential solution named as particles fly through the problem space by following the present optimum particles. Each particle maintains the record of its coordinates in the problem space which are related with the fitness of the particle. This value is referred to as pbest If a particle considers all the population as its topological neighbors the best value is called global best which is also referred to as gbest.

The particle swarm optimization concept consists of steps for changing the velocity towards the pbest and the lbest locations. Acceleration is weighted by a random term, with random numbers which was generated for acceleration towards the pbest and lbest locations. PSO is initialized by a group of random particles. It also searches for optima by updating generations. In Each Iteration, the values are updated using the two values such as pbest and lbest.

The first value is the best value it has ever achieved and it is stored in the memory. It was named as pbest. Another best value tracked by the particle swarm optimizer, the location is called the gbest. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called lbest.

C. Sample Association Rule:

TABLE.1

SAMPLE ASSOCIATION RULES
'HH'
'HLH'
'HLHH'
'HLHHH'
'HHHHHHHLHH'

Fig. 1: Tabular column for sample association rules

The fitness values calculated in corresponding iterations for ABC, PSO and ABC-PSO were given and compared in table 2.

TABLE.2

ITERATION	FITNESS VALUES		
	ABC-PSO	ABC	PSO
10	55.3306	52.0867	53.1192
20	51.7188	50.2728	43.2605
30	51.5286	51.081	45.4033
40	49.0831	48.9424	44.4308
50	53.1716	51.6944	41.5618

Fig.2.Tabular Column For Comparison Of Fitness Values

From the above table it is clear that our proposed method has high fitness value when compared to the existing methods. It is proven that our proposed method is efficient than the existing methods. The fitness values in the y-axis with the corresponding iterations in the x-axis were plotted in the graph. The graph was plotted separately for ABC, PSO, ABC-PSO and for the comparison of all the three methods.

IV. CONCLUSION

The ABC is a new search algorithm. Many approach like ant colonies have been successfully used in Data Mining. It is a challenging task for mining the comprehensive and informative knowledge in such complex data by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining. To extract knowledge from different datasets, we have used a hybrid mining technique. The knowledge extraction was done by association rule mining with the combination of Artificial Bee Colony optimization algorithm (ABC) and Particle swarm optimization (PSO). Hybridization was used to extract the optimal rules from the association rules for further classification. The accuracy was checked in terms of optimal rule. The best position for moving the particle was updated by using ABC algorithm. In future work, we plan to compare the performance of the proposed ABC and PSO algorithm with other algorithms we can also plan to discover a new effort, by changing the parameter values of ABC and PSO.

REFERENCES

[1] Chin-Ang Wua, Wen-Yang Lin, Chang-Long Jiang, and Chuan-Chun Wu, “Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining”,

[2] Sumana Sharma, Kweku-Muata Osei-Bryson, George M. Kasper, “Evaluation of an integrated Knowledge Discovery and Data Mining process model”, Expert Systems with Applications,

[3] Tahar Mehenni and Abdelouahab Moussaoui, “Data mining from multiple heterogeneous relational databases using decision tree classification”,

[4] Haitao Gan, NongSang, RuiHuang, XiaojunTong and ZhipingDan, “Using clustering analysis to improve semi-supervised classification”, Neurocomputing, Vol. 101, pp. 290–298, 2013.

[5] Lee, S.J and Siau, K “A Review of Data Mining Techniques,” Industrial Management & Data Systems, vol.101,no.1,pp. 41-46,2001.

[6] T. Balasubramanian and R. Umarani, “An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique”, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, **(2012)** March 21-23

[7] J. Escudero, J. P. Zajicek and E. Ifeachor, “Early Detection and Characterization of Alzheimer’s Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means”, 33rd Annual International Conference of the IEEE EMBS Boston, assachusetts USA, **(2011)** August 30-September 3.

[8] H. Chipman and R. Tibshirani, “Hybrid hierarchical clustering with applications to microarray data”, Biostatistics, vol. 7, no. 2, **(2009)**, pp. 286-301.

[9] A. Rajkumar and G. S. Reena, Diagnosis of Heart Disease Using Datamining Algorithm”, Global Journal of Computer Science and Technology, vol. 10, no. 10, **(2010)**.