A Multi-Objective Association Rule Mining With Hybrid Approach

Yamee Patel[1], Prof. Maitrey Patel[2]
[1, 2] Grow More Faculty of Engineering, Himmatnagar

Abstract- *Association Rule Mining (ARM) is one of the most well-liked techniques of data mining strategies whose primary aim is to extract associations rule among sets of items or products in transactional databases.Association Rule Mining involves a multi-objective perspective to get an interesting and Precise rule set. By taking into account the Pareto optimality, an optimal trade-off is established between the conflicting and inconsistent performance parameters -comprehensibility, interestingness and confidence of the mined rules. We propose an association rule mining scheme using our proposed multi-objective algorithm using both, Non-dominated Sorting Genetic Algorithm (NSGA-III) and Particle Swarm Optimisation (PSO), being population-based amorphous search method, have found their strong base in mining association rules. The main benefit of the proposed algorithm is that the hybridisation of multiple objective-GA with multi objective-PSO reducing the challenge of choosing one optimization algorithm to solve complex problems also balances the exploration and exploitation tasks, resulting in valuable extraction of accurate and interpretable mined rules.*

Keywords- Association Rule Mining; Multi-Objective Optimization; Pareto optimality; Hybridisation; Genetic Algorithm; Particle Swarm Optimisation; Hybrid NSGA-III - MOPSO;

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amount of data [2]. In Data mining, user can perform analysis of data from different dimensions, categorize it, and recapitulate the relationships identified. Data mining is also considered as synonym of “Knowledge mining from data”. The process of finding knowledge from large amount of data is known as Knowledge Discovery from Data (KDD). In data mining, Association rule mining is used to uncover an association among different items from the analysis of a large dataset [13]. To find association rules, two important measures minimum support and minimum confidence are used. Both minimum support and minimum confidence are user-specified measures. Association rule mining is the process of two steps. In first step frequent item sets are generated and in second step rules are discovered. Association rules are like, when coffeeis purchased; sugar is purchased 80% of the time. The second one states that 60% of the time when bread is purchased so is jelly. Association rules are generally used in market-basket analysis. Because it is used to know buying behavior of customer to get more profit. Mining association rules from large transactional database has been considered as an significant part of database research. Association rule mining is used to find relationships between items and generated results can provide a basis for decision making and forecasting[9].Genetic algorithm is a method of heuristic. Heuristic method is generally used to produce helpful solutions for optimization. Genetic algorithm is part of the bigger class of evolutionary algorithms. This algorithm generates solution to optimization problem. Genetic Algorithm is a part of computational model[9]. For discovering optimized association rules, genetic algorithm plays an importantrole.Becausegeneticalgorithmperformsglobalsearchf ordiscovering frequency of set of items. Complexity of genetic algorithm is less than other algorithms used for mining important data. To discover association rules genetic algorithm has been carried out which is also used in problems like biology, fraud detection and commercial databases.

NSGA-III approach has been applied to three to 15-objective existing and new test problems and to three and nine-objective practical problems. The test problems involve fronts that have convex, concave, disjointed, differently scaled, biased density of points across the front, and multimodality involving multiple local fronts where an optimization algorithm can get stuck to. In all such problems, the proposed NSGA-III approach has been able to successfully find a well-converged and well diversified set of points repeatedly over multiple runs. The performance scaling to 15-objectives is achieved mainly due to the aid in diversity preservation by supplying a set of well distributed reference points. Here, they replace the crowding distance operator with the following approaches:-

A. Classification of Population into Non-dominated Levels
B. Determination of Reference Points on a Hyper-Plane
C. Adaptive Normalization
D. Association Operation
E. Niche-Preservation Operation

Multi-objective problem PSO has significantly used, in which the objective function domination amonf. In this paper External Repository stores all the non-dominated solutions obtained so far along the search progress. After randomly generating the swarm particles, domination among particles is determined and the External Repository is initialised with current non-dominated swarm particles. At each Generation, for each Swarm Particle, by using Roulette Wheel Selection, a leader, the particle that guides other particles towards better regions of the search space, is selected from the repository. In accordance with the leader's parameters, the particle's flight is changed, which is followed by optional mutation. After evaluation, Pbest (Best personal position of each particle) is updated to maintain each particle's best position. After each swarm particle in a generation is updated, non-dominated swarm particles are identified to update the limited size repository. At the end of all the generations, the repository containing all the non-dominated solutions is returned.

II. LITERATURE SURVEY

A number of Multi-Objective Evolutionary Algorithms have been proposed in the literature for ARM, one of the earliest is by the authors Ghosh and Nath [12]. The authors [12] used GA alongwith Pareto Optimality for multi-objective optimization. One of the drawbacks of their approach was that only a sample of data was used to calculate support count of various attributes. In [13], the authors have employed Non-dominated Sorting Genetic Algorithm II (NSGA-II) as an association rule miner for a multi-objective problem. Two pitfalls of conventional Multi-Objective Evolutionary Algorithms (MOEA) have been listed. First, good rules generated during intermediate generations are lost. Second, non-viable rules, such as empty consequent, are not eliminated. In [1], the authors have developed Binary PSO for association rule mining, with fitness value being calculated on basis of support and confidence. Although, the end user does not need to specify the threshold values for implementing this approach, but it is required for the user to mention a number of rules, that he/she wants to be generated from the given dataset. In [14], the authors have tried to incorporate evolutionary learning by proposing an extended Multi-Objective Evolutionary algorithm. External population has also been introduced to mine reduced set of positive and negative Quantitative Association Rule (QAR). In [9], using confidence and support as the only parameters, the authors have employed hybridisation of GA and PSO for mining of association rules. Although exploration and exploitation tasks have been balanced but the problem with this approach is that they do not provide high quality rules, useful for the user. Further, we have extended this methodology to incorporate multi-objective optimization technique by involving – performance, comprehensibility and interestingness.

III. MULTI-OBJECTIVE ASSOCIATION RULE MINING

- Association rule mining problem is a multi-objective problem not a single objective one[16]. Various objectives of association rule mining problem are confidence, interestingness and comprehensibility. By using measures like confidence, interestingness and comprehensibility we can get easily understandable rules.
- Confidence: Confidence factor or predictive accuracy of a rule is defined as[16]:

Confidence (X→Y) = Support (X∪Y) / Support (X)

- Comprehensibility: It is calculated by the number of attributes used in antecedent part of the rule relating to the consequent part of the rule.

Comprehensibility of a rule X ⇒ Y is measured by:

Comprehensibility = log (1+ |C|) / log (1+ |AUC|)

Where, |C| is the number of attributes present in the consequent part and |AUC| is the number of attributes involved in the total rule.

- Interestingness: It is used to measure how much the rule is astonishing for the user.

Interestingness = [SUP(AUC)/SUP(A)]
*[SUP(AUC)/SUP(C)] * [1(SUP(AUC)/|D|)]

Where, |D| represents the total number of records in the database.

IV. PROPOSED METHODOLOGY

Most of the traditional methods for association rule mining (ARM) use 'Support' and 'Confidence' measures for rule mining. Finding Frequent itemset that is the first this in extraction of association rule. Rules have to satisfy minimum support if not then eliminated. In the second phase, rule have to satisfy confidence if not then eliminated. Support values define how many rules truely existed in dataset. Confidence measure is:

Confidence = SUP (AUC) / SUP (A)

Association rules are usually shown with A→C, where 'A' stands for antecedent part of the rule and 'C' stands for the consequent part. This simply means that if A is present, then C will be present as well. But this measure cannot guarantee obtaining suitable association rules individually. In addition to having appropriate coverage and reliability, generated rules should also be interesting and comprehensible.

This manner, the problem of ARM becomes a multi-objective problem instead of a single-objective one.

Some researchers have characterized this as an optimization problem and tried to mine ARS using global optimization algorithms. Genetic algorithms (GA) and Particle Swarm Optimisation (PSO) are one of the best global optimization algorithms and because of our problem's nature – having multiple objectives in Association rule mining; a multi-objective genetic algorithm can be a useful approach. So we are taking hybridization approach and combining two population based naturally inspired global optimization algorithms: NSGA-III and MOPSO. The system main aims to output high quality rules.

Algorithm:

1) Initialize Population
2) Generate random N-size Population
3) $F = f(\hat{x}) \leq f(x), \forall x \in Rn$ (The function f is called the objective function, which maps the search space to the function space., which will provide a single fitness value for each parameter)
4) sort solutions
    a) Group by ascending order of front Rank based on each member with reference point (key -> value pair)
5) **for each** Generation in Rank based matrix **do**
6) Create offspring population
    a) Roulette wheel selection, crossover, mutation
        i) Calculate the sum of all fitness's in population (sum S).
        ii) Generate a random number r in the interval [0; S].
        iii) Go through the population and sum fitness's. When the sum s is greater than r, stop and return the individual where you are.
7) Combine Parent and Offspring Population
8) For each particle ,do
    a) Select Leader on Random basis
    b) Update particle position and velocity
    c) Evaluate Objective function(Same as step 3)
    d) Update current fitness
9) Sort and Select solutions based on latest fitness found from matrix
10) Group by ascending order of front fitness rank
11) Associate each member with reference point
12) Display in descending order

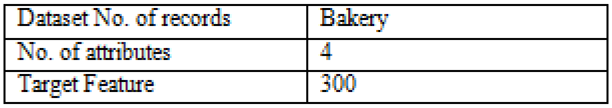
Hybridisation, a burgeoning field in Computational Intelligence, goals to integrate desirable features of different algorithms, while curbing their individual flaws. GA and PSO, both are stochastic population based approaches, hence their combination has been a successful approach. The technique to create new reliable solutions, in other words, the mechanism to explore the search space, differs in these algorithms. In GA, the chromosomes breed with each other to produce off-springs whereas, in PSO, particle positions are influenced by their self-learned knowledge as well as by the information sharing among swarm members. While PSO has a fast convergence compared to GA operators - mutation and crossover, feasibility of real-world problems based heuristics can be successfully incorporated in GA using these same conventional operators – mutation and crossover for feasible solutions, making such unlikely for PSO approach, Thus, unlike the many proposed algorithms, we intend to integrate both global optimisation approaches – GA and PSO for our multi-criteria problem.

In our Proposed methodology first we initialize population, then generating random N- size Population, For random N -size Population we calculating Objective Function In Which for each Parameter Objective Function like Confidence, Comprehensibility, Interestingness are evaluated based On That single fitness value for Each Parameter are calculated. After calculating fitness for each parameter sorting is performed, in which assign reference point to the each of the member and grouping into ascending order of the rank based on fitness value. For each Generation in rank based matrix we are generating offspring population, using Roulette wheel selection method offspring population are generated.This selection method is based on the concept of proportionality. Parents are selected based on a probability value. In this selection, the opportunity to obtain a good solution is quite high. Conceptually, the fitness value of each individual or potential solution in a population corresponds to the area of the roulette wheel proportions. When the roulette wheel is spun, a solution marked by the roulette wheel pointer is then selected. The higher fitness value with a bigger area is likely to have more chances of being chosen. The segment size and selection probability remain the same throughout the selection phase. The advantage of this technique is that it gives no bias with unlimited spread. However, one of its disadvantages is that it cannot handle negative fitness values due to the proportionality concept. In addition, it could not handle a minimization problem directly, but this limitation can be overcome by transforming it into an equivalent maximization problem. Then combine parent and offspring population. For each Member select the leader which help to update particle position, after that objective function is evaluated using the result of objective function update fitness of particle. Non-domination sorting is performed during this stage, in which similar fitness having members are combined into the group, and non-dominated groups are discarded.Dominant members are group by ascending order of fitness rank. Then Assign reference point to each member of the group and display frequent itemsets.

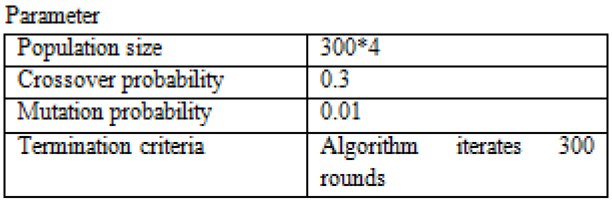
V. RESULTS

The Proposed algorithm applied to three datasets following are:-

Properties of test datasets:

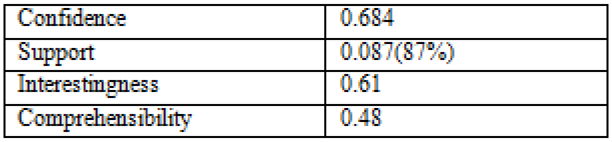


Dataset No. of records	Bakery
No. of attributes	4
Target Feature	300



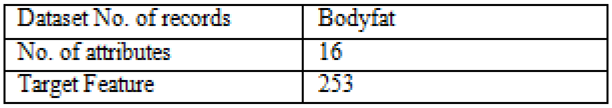
Parameter

Population size	300*4
Crossover probability	0.3
Mutation probability	0.01
Termination criteria	Algorithm iterates 300 rounds

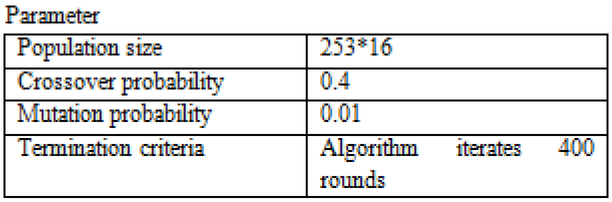


Confidence	0.684
Support	0.087(87%)
Interestingness	0.61
Comprehensibility	0.48

Properties of test datasets:

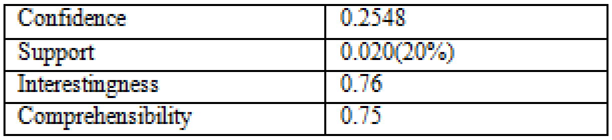


Dataset No. of records	Bodyfat
No. of attributes	16
Target Feature	253



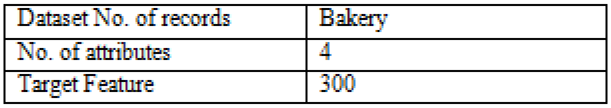
Parameter

Population size	253*16
Crossover probability	0.4
Mutation probability	0.01
Termination criteria	Algorithm iterates 400 rounds

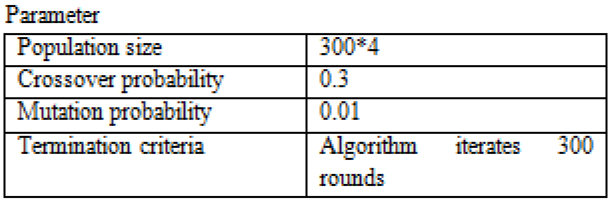


Confidence	0.2548
Support	0.020(20%)
Interestingness	0.76
Comprehensibility	0.75

Properties of test datasets:

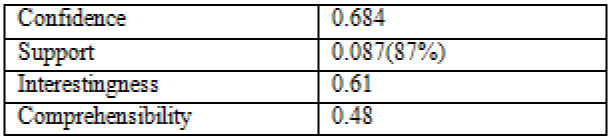


Dataset No. of records	Bakery
No. of attributes	4
Target Feature	300



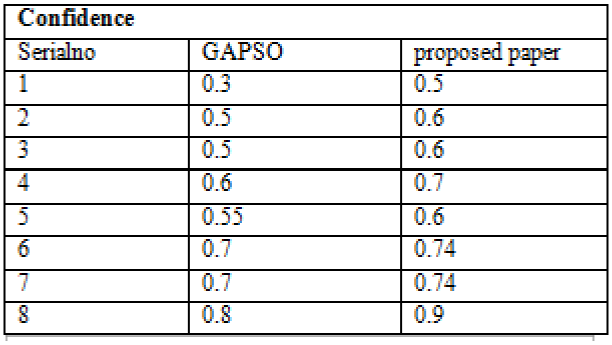
Parameter

Population size	300*4
Crossover probability	0.3
Mutation probability	0.01
Termination criteria	Algorithm iterates 300 rounds

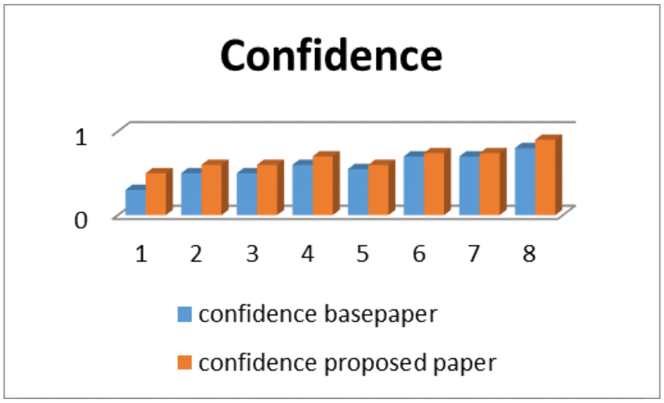


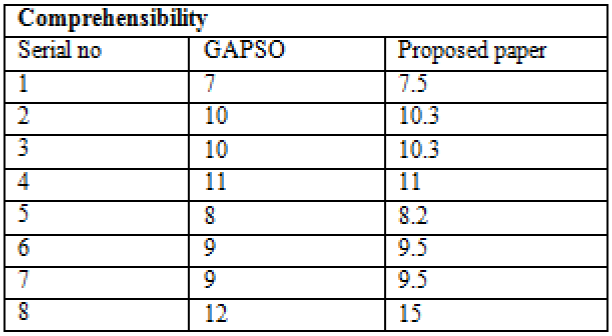
Confidence	0.684
Support	0.087(87%)
Interestingness	0.61
Comprehensibility	0.48

Comparison Result:

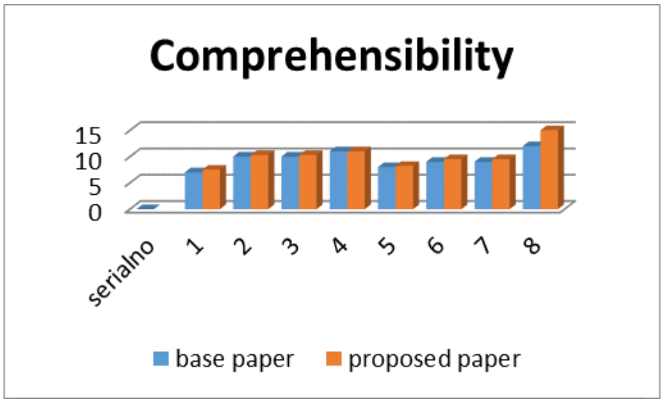


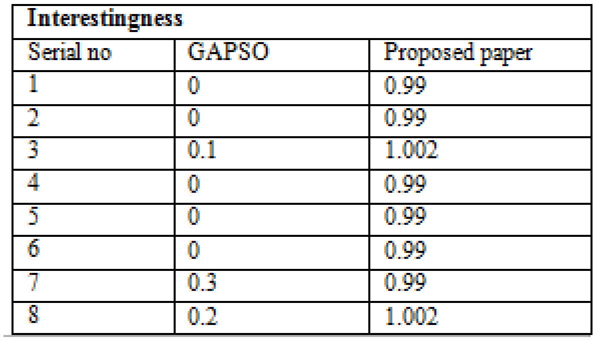
Confidence		
Serialno	GAPSO	proposed paper
1	0.3	0.5
2	0.5	0.6
3	0.5	0.6
4	0.6	0.7
5	0.55	0.6
6	0.7	0.74
7	0.7	0.74
8	0.8	0.9



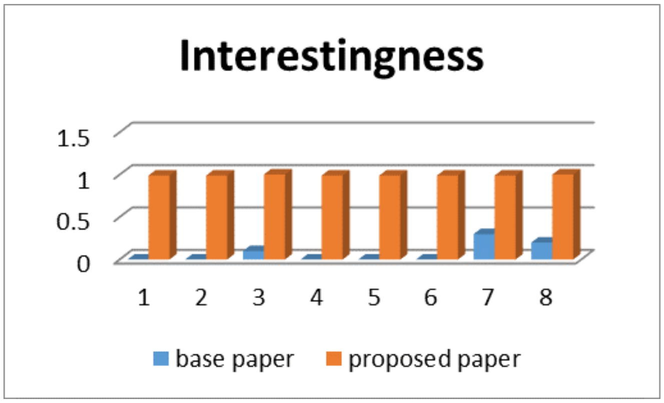


Comprehensibility		
Serial no	GAPSO	Proposed paper
1	7	7.5
2	10	10.3
3	10	10.3
4	11	11
5	8	8.2
6	9	9.5
7	9	9.5
8	12	15





Interestingness		
Serial no	GAPSO	Proposed paper
1	0	0.99
2	0	0.99
3	0.1	1.002
4	0	0.99
5	0	0.99
6	0	0.99
7	0.3	0.99
8	0.2	1.002



VI. CONCLUSION

In this Approach, we have proposed a new multi-objective optimization model as an association rule miner, which is based on hybrid algorithm of the two population based naturally inspired global optimization algorithms: NSGA-III and MOPSO. The model aim is to extract high quality rules. To employ the principle of Multi-Objective Optimisation, Pareto Optimality has been used. NSGA-III approach has been able to successfully find a well-converged and well diversified set of points repeatedly over multiple runs. In PSO, particle positions are influenced by their self-learned knowledge as well as by the information sharing among swarm members. In proposed methodology bakery dataset used, and used to reduce execution time of the algorithm.

REFERENCES

[1] Wilson Soto and Amparo Olaya–Benavides" A Genetic Algorithm for Discovery of Association Rules" IEEE 2011.

[2] Basheer M. Al-Maqaleh , Hamid Shahbazkia, "A Genetic Algorithm for Discovering Classification Rules in Data Mining ," IJCA March 2012.

[3] Arvind Jaiswal, "Generalized and Identify the Best Association Rules using Genetic algorithm," IJSR, Volume 3 Issue 6, June 2014.

[4] Peter P. Wakabi-Waiswa and Venansius Baryamureeba, "Mining High Quality Association Rules Using Genetic Algorithms,".

[5] Mrinalini Rana and P S Mann, "Association Rule Mining with Multi-Fitness Function Genetic Algorithm," IJSETT, May 2013 .

[6] Rupali Haldulakar, Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm," IJCSE, Vol. 3, No. 3, Mar 2011.

[7] Mohit K. Gupta and Geeta Sikka , "Association Rules Extraction using Multi-objective Feature of Genetic Algorithm ," WCECS 2013.

[8] B. Minaei-Bidgoli, R. Barmaki, M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms," Elsevier Inc., 2013.

[9] Dimple S. Kanani , Shailendra K. Mishra , "An Optimized Association Rule Mining using Genetic Algorithm," IJCA, Volume 119 – No.14, June 2015 .

[10] Ye Gao, Zhe Liu, "Mining Association Rules with Constraints Based on Immune Genetic Algorithm," IEEE, 2015.

[11] N. Srinivas and K. Deb, "Multiobjective function optimization using nondominated sorting genetic algorithms," Evol. Comput., vol. 2, no. 3, pp. 221–248, Fall 1995.

[12] Ghosh, Ashish, and Bhabesh Nath. "Multi-objective rule mining using genetic algorithms." Information Sciences 163.1 (2004): 123-133.

[13] Anand, Rajul, Abhishek Vaid, and Pramod Kumar Singh. "Association rule mining using multi-objective evolutionary algorithms: Strengths and challenges." Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, 2009.

[14] Martin, D., Rosete, A., Alcalá-Fdez, J., & Herrera, F. (2014). A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. IEEE Transactions on Evolutionary Computation, 18(1), 54-69.

[15] Kalyanmoy Deb, Associate Member, IEEE, Amrit Pratap, Sameer Agarwal, and T. Meyarivan, "A Fast and Elitist

Multi objective Genetic Algorithm:NSGA-II," IEEE, 2002.

[16] Kalyanmoy Deb, Fellow, IEEE and Himanshu Jain," An Evolutionary Many-Objective Optimization Algorithm Using Reference-point Based Non-dominated Sorting Approach, Part I: Solving Problems with Box Constraints," IEEE 2013.

[17] Coello, CA Coello, and Maximino Salazar Lechuga. "MOPSO: A proposal for multiple objective particle swarm optimization." Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on. Vol. 2. IEEE, 2002.

[18] Aashna Agarwal, Prof(Dr.) Nirali Nanavati, "Association rule mining using hybrid GA-PSO for multi-objective optimisation," IEEE, 2016.

Book

[1] "Data Mining, Concepts and Techniques," Jiawei Han, Michline Kamber, Jian Pei.