International Journal of Mechanical Engineering and Technology (IJMET)
Volume 8, Issue 11, November 2017, pp. 999–1008, Article ID: IJMET_08_11_102
Available online at http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=8&IType=11
ISSN Print: 0976-6340 and ISSN Online: 0976-6359

© IAEME Publication **Scopus** Indexed

ASSOCIATION RULE MINING USING BIO-INSPIRED BEES AND SWARM INTELLIGENCE ALGORITHMS

D. Ravikiran
Research Scholar Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, A.P, India

Dr S.V.N Srinivasu
Research Supervisor, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, A.P, India

ABSTRACT:

Association Rule Mining (ARM) is sound considered an important optimization problem which discovers worthwhile rules from given transactional databases. Several algorithms suggested in the literature which confirmations their adeptness when distributing with different sizes of datasets. Inappropriately, their adeptness is not adequate for conduct large-scale datasets. Bio-inspired bees swarm intelligence algorithm for association rule mining is more efficient. These kinds of problems need more powerful processors and are time expensive. For such issues solution can be provided by graphics processing units (GPUs). They are massively multithreaded processors. In this case, GPUs can be used to increase the speed of the computation. Bees and swarm intelligence algorithm for association rule mining can be designed using GPUs in the multithreaded environment which will efficient for given datasets. The proposed system is designed and developed with following modules; we propose two parallel versions of the BSO-ARM algorithm on GPU architecture called SE-GPU and ME-GPU algorithms standing for single evaluation on GPU and multiple evaluations on GPU, respectively.

Keywords: Association Rule Mining, CUDA, GPUs, Bio-inspired, Swarm Intelligence.

Cite this Article: D. Ravikiran and Dr S.V.N Srinivasu, Association Rule Mining Using Bio-Inspired Bees and Swarm Intelligence Algorithms, International Journal of Mechanical Engineering and Technology 8(11), 2017, pp. 999–1008.
http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=8&IType=11

1. INTRODUCTION:

Swarm Intelligence and bio-inspired computation have become increasingly popular in the last two decades. Bio-inspired algorithms such as ant colony algorithms, bat algorithms, bee algorithms, firefly algorithms, cuckoo search and particle swarm optimization have been applied in almost every area of science and engineering with a dramatic increase of a number of relevant publications.

Solving combinatorial optimization problems is time-consuming when dealing with large-scale data. Metaheuristicshave been largely used to solve this kind of problems. They can be viewed as general purpose approaches based on stochastic methods, exploring outsized search spaces to find near-optimal solutions in a reasonable time. Some metaheuristics are inspired by biological and physical phenomenon. During the last two decades, two population-based methods such as evolutionary algorithms (EA) and swarm intelligence (SI) showed their efficiency compared to other metaheuristics.

Association Rule Mining (ARM) is one of the most useful and well-known techniques of data mining [1]. It is used to extracts or retrieves frequent patterns, associations, correlations, or causal structures among sets of items from given datasets. Datasets can be considered as transactional databases, relational databases, and other information repositories. Formally, association rule mining problem explained as follows: Let, T be m number of transactions or set of records like {t1, t2, t3,……, tm} from given datasets. And, I be a set of n different items or attributes {i1, i2, i3 ,……..…, in}, an association rule is an implication of form $X \rightarrow Y$ where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The itemset X is called antecedent part (left side) while the itemset Y is called consequent part (right side) and the rule means X implies Y.

Association Rule Mining is related to finding a set of rules considering a large percentage of data, and it tends to generate a useful number of rules. However, since the number of transactions is increasingly more, the user no longer looks for all the possible rules, but user looks to determine only a subset of important rules. To measure the usefulness of association rules, mainly two basic parameters are used, one is namely support of a rule and second is confidence of rule. The support of an itemset $I' \subseteq I$ is the number of records containing I′. The support of rule $X \rightarrow Y$ is the support of $X \cup Y$ and the confidence of a rule is support ($X \cup Y$) / support (X). An association rule $X \rightarrow Y$ with a confidence of 70% means that 70% of the transactions that contain X also contains Y together. So, the association rule mining is to find important rules which having support ≥ MinSup and confidence ≥ MinConf [2]. Here, MinSup and MinConf are two thresholds predefined by users.

Swarm intelligence is based on the collective behavior of decentralized, self-organized systems. Many researchers are interested in this new existing way of achieving a form of artificial intelligence including simple agent groups. Modelling the behavior of social insects such as ants, bees, firefly, bat etc. and using these models for search and problem-solving are the context of the emerging area of swarm intelligence. Bees are among the well-studied social insects [3]. Many researchers studied the bees behavior to design powerful methods.

The bees behavior is divided into three categories: marriage behavior, foraging behavior, and the queen bee. Marriage bees behavior [4] approach, is used in honey bees optimization. Bees are social insects living in organized colonies. Each honey-bees colony consists of one or several queens, drones, workers, and broods. Queens specialize in egg laying, workers in brood care and sometimes egg lying, drones are the males of the colony and broods the children.

The meta-heuristic BSO proposed in [5] is inspired by the foraging bees behavior. It is based on a swarm of artificial bees cooperating to solve a problem. The general functioning of this is as follows: First, a bee named Initial Bee settles to find a solution presenting good

features. From this first solution called Sref, we determine a set of other solutions of the search space by using a certain strategy. This set of solutions is called Search Area. Then, every bee will consider a solution from Search Area as its starting point in the search. After accomplishing its search, every bee communicates the best-visited solution to all its neighbors through a table named dance. One of the best solutions stored in this table will become the new reference solution during the next iteration. To avoid cycles, the reference solution is stored every time in a taboo list. The reference solution is chosen according to the quality criterion. However, if after a period the swarm observes that the solution is not improved, it introduces a criterion of diversification preventing it from being trapped in a local optimum. The diversification criterion consists to select among the solutions stored in the taboo list, the most distant one. The algorithm stops when the optimal solution is found.

Nowadays Graphics processing units (GPUs) is fast and cheap parallel hardware traditionally used to speed up 3D graphic applications. GPU-accelerated computing is the use of a graphics processing unit (GPU) together with a CPU to accelerate deep learning, data analytics, and engineering applications. It is pioneered in 2007 by NVIDIA, GPU accelerators now power energy-efficient data centers in government labs, universities, enterprises, and small-and-medium businesses around the world. They play a huge role in accelerating applications in platforms ranging from artificial intelligence to cars, drones, and robots. This new technology general purpose graphics processing units also known as GPGPU. GPU-accelerated computing offloads compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. From a user's point of view, applications simply run much faster. Hence it used to improve the execution time of various computation from different domains.

Motivating by graphics processing units (GPUs), the power of GPUs can be used to solve highly computational problems from computer science domain. So, a bees swarm optimization for association rule mining using GPUs is useful approach than the serial approach of Bees swarm Intelligence. The evaluation process of the solutions could be done on GPU because of GPUs massively multithreaded environment.

2. RELATED WORK:

Many sequential algorithms have been designed for an ARM problem. Some algorithms are based on exact approaches like Apriori [7], FPgrowth [8], DIC [7], and DHP [38]. Apriori and FPgrowth are the most used algorithms. Nevertheless, these algorithms are time intensive when applied to large datasets.

Some well-known algorithms have been proposed for generating association rules are AIS [6], Apriori [7] and FP-Growth [8].

Agrawal R, I mielinski T and Swami A,[6] have been proposed AIS algorithm which is very space consuming and requires too many passes over the whole database. Agrawal, R. and Ramakrishan, S [7] has given Apriori algorithm which is the best well-known algorithm for association rules mining. It is basically based on breadth-first search (BFS) strategy to count the supports of itemsets and uses a candidate generation function to achieve the downward closure property of support. Han, J., Pei, J., Yin, J., Mai, R. [8], have been proposed FP-growth algorithm. It uses an FP-tree construction to wrapping the database and a divide-and-conquer approach, to decompose the mining tasks and the database as well. But tree generation is always complex part of data organization.

Agrawal and Shafer [9] projected two parallel versions of Apriori called count distribution (CD) and data distribution (DD). These are leading parallel ARM algorithms. In CD the dataset is divided in between several processors, and each processor executes the entire Apriori on its part of data. Beyond all these algorithms, different parallel metaheuristics have

been proposed in the literature for the ARM problem. Melab N, Talbi E-G [10] proposed a parallel genetic algorithm (GA), called PGARM, based on the master/workers model.

Following this brief state of the art, different ways have been proposed for the parallel ARM problem. For each approach, the authors take advantage of the used parallel hardware. However, all these algorithms have been implemented on old-fashioned parallel architectures which are still expensive and not always available for everyone.

Since 2007, GPU technology, fast, cheap, and parallel hardware, usually available on most computers, has been successfully used in many domains. For this, ARM community is also investigating on this simple but reliable technology. To the best of our knowledge, the only work which introduces metaheuristics for ARM on GPU is proposed in [11] by Cano et al. In this paper, an evolutionary algorithm is proposed to solve the ARM problem on GPUs.

Djenouri Y, Drias H, Habbas Z in [12], [13] applied Bees algorithm for web association rule mining, and bees swarm optimisation using multiple strategies for association rule mining respectively. In [14] recent Youcef Djenouri, Habbas proposed the algorithm for GPU based bees swarm optimization for association rule mining.

3. METHODOLOGY

Swarm Intelligence

Swarm intelligence is an emerging field of biologically-inspired artificial intelligence based on the behavioral models of social insects such as ants, bees, wasps, termites, etc

Swarm Intelligence is:

- One Million Heads, One Beautiful Mind
- Agents interacting locally with each other and the environment
- Agents follow simple rules
- Emergence of Intelligent, Collective, Self-organised, Global behavior
- Decentralized and artificial or natural
- Very adaptive
- Randomness enables the continuous exploration of the alternatives, and it ensures that the better solution will be found.
- Behavior relies on stochastic choices made by the agents which are a balance between a simple perception-reaction model and a random model.
- Application of bio-inspired concepts
- A Large mass of the agents is a must.

Swarm Languages:

1) It is a true distributed programming language.
2) The Fundamental Concept behind Swarm: We should move the computation, not the data.
3) The Swarm Prototype: It is a stack-based language, similar to a primitive version of the Java bytecode interpreter and is now implemented as a Scala library.

3.1. ParallelApproaches based on GPU:

The proposed system is designed and developed with following modules; we propose two parallel versions of BSO-ARM algorithm on GPUarchitecture called SE-GPU and ME-GPU algorithms standing for single evaluationon GPU and multiple evaluations on GPU, respectively.

3.2 Parallel ARM approaches based on GPU.

SE-GPU algorithm

The SE-GPU algorithm is based on the master/slave paradigm. The master is executed on CPU, and the slave is offloaded to the GPU. First, the master initializes randomly

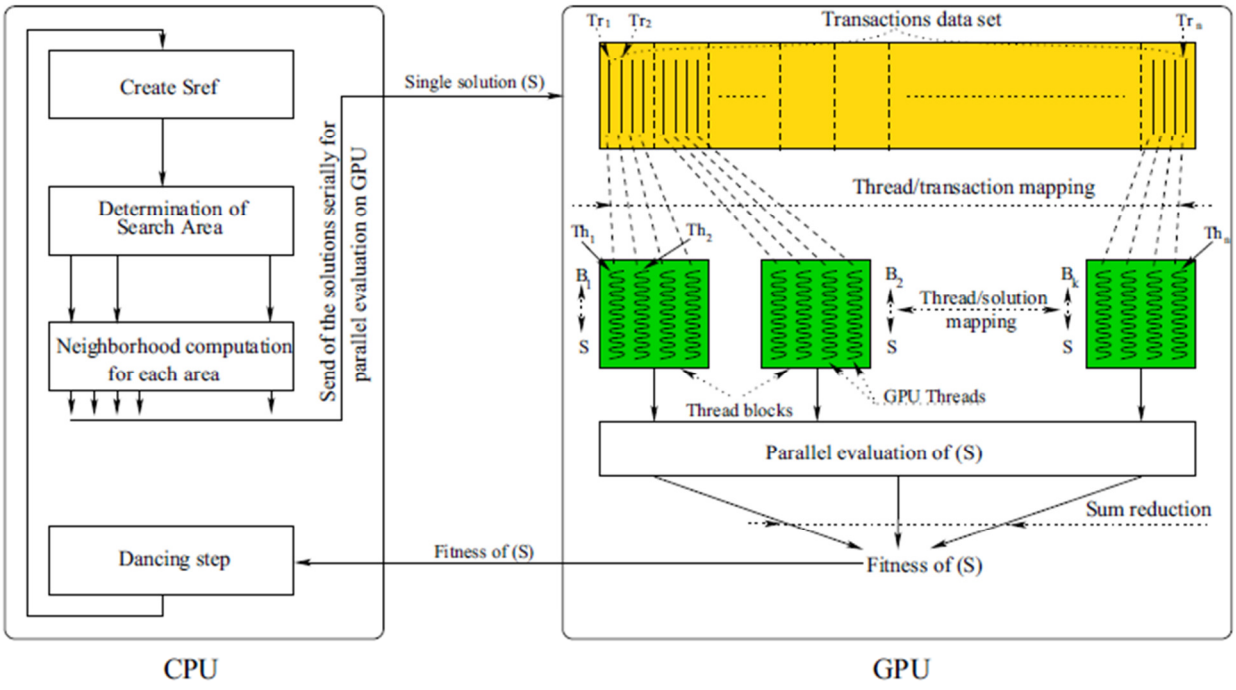


Figure 1 Bees Swarm Intelligence for Association Rule Mining on CUDA

The mentioned module details were given as follows.

3.1.1. Create Initial Solution

Generate initial solution randomly for considering N items. Solution representation is important consideration in this project. Integer encoding representation allows separating the antecedent part and the consequent part of the association rule. And, calculate the cost of the initial solution which is based on support and confidence of the given solution. The cost is known regarding system is fitness for given solution. The rule is considered as one solution in the search space, each one is represented by a vector S of N bits, and their positions are defined as follows:

- S[i] = 0 if the item i is not in the solution S.
- S[i] = 1 if the item i belongs to the antecedent part of the solution S.
- S[i] = 2 if the item i belongs to the consequent part of the solution S.

Example:

Let T: {t1, t2, ……. , t5} be a set of items

S1: {1, 1, 0, 0, 2} represents the rule

R1: t1, t2 ⇒ t5.

Below Fig. 2 shows detail about how to represent rule. (Integer Representation). For given rule only three items which are considered Bread, Peanuts, and Jam. Bread and Peanuts are consequent part of the rule and Jam is consequent part of the rule.

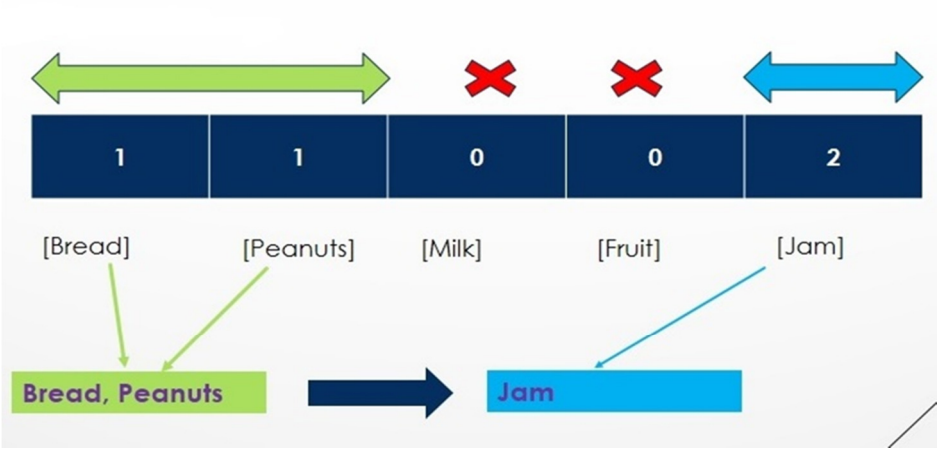


Figure 2 Integer Encoding Rule Representation

3.1.2 Search Area Determination

Given an Initial reference solution, Sref and a colony of K bees the search area operations determine K search spaces; each one is associated to a bee.

Each bee k builds its own search area by changing successively in the solution Sref the bits k + i × Flip where i varies from 0 to n − 1 and Flip is a given parameter. This strategy can be used if and only if the number of bees is less or equal to N/Flip.

For the search area, the aim is to determine the regions of the bees using Sref already created. The strategies have been developed to explore search area. For example, the strategy aims to perform the flip jump on Sref. If we have k=3 and flip=2 and N=5,

We obtain:

- the first bee is obtained by modifying the bits (1, 3, 5)
- the second bee is obtained by modifying the bits (2, 4) and
- the last bee is obtained by modifying the bits (3, 5)

3.3.3. Neighborhood Computation

The neighborhood computation for each search area is obtained by changing from a given solution S one bit in a random way. Based on this simple operation, N neighborhoods are created

Example:

Consider the given solution: S = {1, 0, 0, 1, 2}

1. Change the first bit in S: S1 = {0, 0, 0, 1, 2}
2. Change the second bit in S: S2 = {1, 2, 0, 1, 2}
3. Change the third bit in S: S3 = {1, 0, 1, 1, 2}
4. Change the fourth bit S: S4 = {1, 0, 0, 0, 2}
5. Change the fifth bit S: S5 = {1, 0, 0, 1, 0}

All neighbors send serially to evaluate fitness of the solution.

3.3.4. Fitness

In this Module for each generated solution (rule) from neighborhood computation, the entire transactional database is scanned. The solution fitness is based on the support and the confidence of the given rule which is computed as follows:

$$\text{Fitness}(s) = \alpha \times \text{confidence}(s) + \beta \times \text{support}(s) \quad (3.1)$$

This function should be maximized. For each invalid solution s where Sup(s) < Minsup or Conf(s) < MinConf, the Fitness(s) is set to −1 and the solution is rejected.

3.3.5. Dance Table

Each bee puts in the dance table the best rule found among its search. The communication between bees is done to find the best dance (the best rule) which becomes the reference solution for the next pass. The general functioning of the algorithm is as follows: First, the initial solution reference (Sref) is initialized arbitrarily so that each element of Sref belongs to {0,1,2}. After that, excluding the Fitness Computing which is applied for each generated solution, the other steps are repeated in the order until maximum iteration is reached.

These five modules combinedly work on using CPU and GPU. Fig.1 shows were working on these modules. CPU runs master model, and GPU runs slave which consists on evaluation fitness for given solution.

ME-GPU algorithm

Like SE-GPU, ME-GPU algorithm is also based on the master/workers paradigm. The master is executed on CPU and the slave as a kernel on GPU. Here, the bees explore their regions on CPU and evaluate these solutions on GPU. Unlike in SE-GPU, in MEGPU several solutions are evaluated simultaneously on GPU. The master generates the rules and then sends them to the GPU. The GPU on its side evaluates them in parallel and sends back the fitness values to the master. This process is repeated as in SE-GPU until the number of maximal iterations is reached.

Evaluating only one solution at a time on GPU allows us to speed up the calculation time of the evaluation process. However, transactional databases are composed of a huge number of transactions generating a huge number of rules to evaluate on GPU. Consequently, a huge amount of data must be communicated between CPU and GPU degrading the overall performance of the Algorithm. Therefore, CPU/GPU communications should be minimized. To do so, the evaluation process is performed on multiple rules at a time on GPU. Indeed, each block of threads is mapped to one rule (see Fig. 3). Threads of the same block are launched to calculate collaboratively the fitness of a single rule. Therefore, there are as many rules as blocks. The transactions are subdivided into subsets, and each subset set is assigned to exactly one thread thi so that each thread calculates only this part of the transactions set. After that, a sum reduction is applied to aggregate the fitness value.

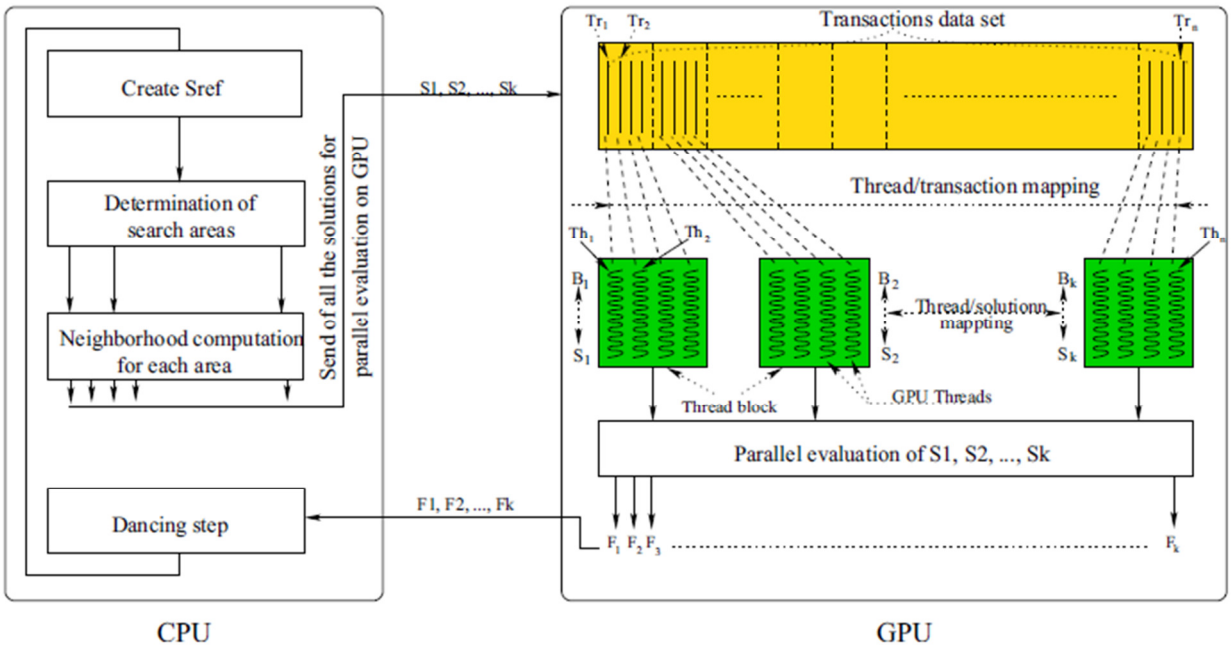


Figure 3 ME-GPU framework

4. HELPFUL HINTS:

Implementation of the Parallel approach of the proposed system is carried out using CUDA enabled GPUs. CUDA is a parallel computing platform and application programming interface (API) model created by Nvidia. It permits software developers and software engineers to use a CUDA-enabled graphics processing unit (GPU) for general purpose processing – an approach termed GPGPU (General-Purpose computing on Graphics Processing Units). The CUDA platform is a software layer that gives uninterrupted access to the GPU's instruction set and parallel computational elements.

Above proposed system is design and developed using general approach of bees swarm intelligence algorithm. Algorithm 4.1 gives general bees algorithm.

Algorithm 4.1: The general Bee Swarm Intelligence algorithm

Input: transactional database.

Output:number of solutions.

Sref ← The solution found by Initial Bee.
    while i < Max_Iteration and not stop do
Insert Sref in the taboo list.
Search-Area(Sref).
Assign a solution from SearchArea to each bee.
For each bee K do
Built-Search-Area (beek).
Store the result in the table Dance.
Communication between bees to choose best solution
end for
Choose the new reference solution Sref.
end while

5. EXPERIMENTAL SETUP AND RESULT

The proposed system is performed using specific software and hardware requirements.

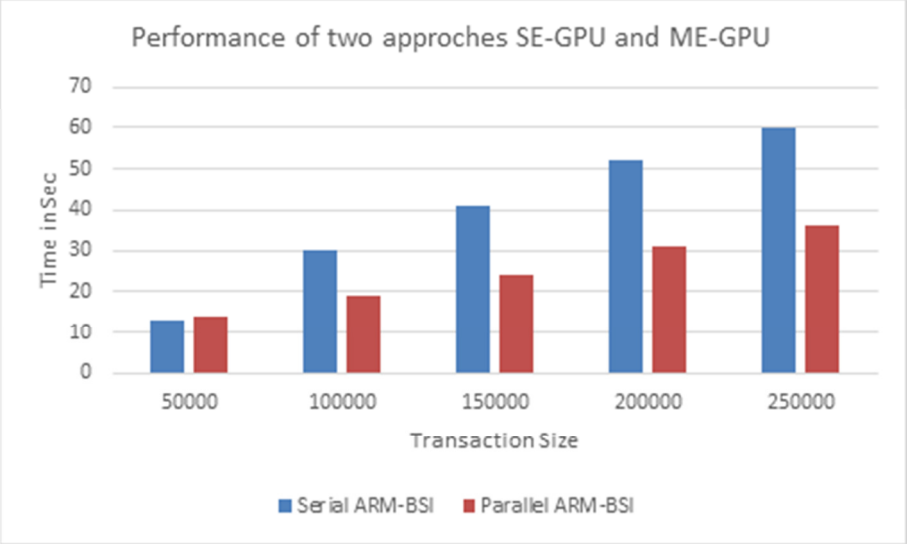
5.1. Software requirements for experiment:

Proposed system is implemented using Windows 10, 64-bit operating System, using C as programming language. Important setup is CUDA 7.0 API**.**

5.2. Hardware requirements for the experiment:

Intel i5 4th generation dual-core processor, Clock Speed 2.7GHz, 4 GB memory and machine which having GT 730 Nvidia graphics card with 386 cores

Proposed System is tested. Figure 4 shows the speedup of the two approaches (SE-GPU and ME-GPU) according to different datasets. Notice that there is a high difference between the two approaches using a condensate data (number of items is very high), and small difference using non-condensate data. These results validate the first experimentation.ME-GPU's speedup exceeds 290, while SE-GPU's speedup does not exceed 100. We can say after this experimentation that if we have large non-condensate data, then SE-GPU is more appropriate by choosing a small number of bees. Otherwise, if we have large and condensate data, then ME-GPU is more appropriate by choosing a big number of bees.



6. CONCLUSION

To validate the effectiveness of the proposed approach, the two algorithms were implemented and tested on various datasets. Extensive experiments have been performed. They show that the speed of the two approaches is about a hundred-time speedup concerning the one of the sequential BSO-ARM algorithm. Moreover, SE-GPU outperforms ME-GPU for the non-condensed datasets

Bio-inspired swarm intelligence based problem approaches are more powerful than traditional problem-solving techniques. If these approaches are applied to solve the highly computational problem like mining problems on large scale dataset, then it is time-consuming task. In that case, GPGPU can be used to divide this task and solve parallel using massively multithreaded environment of GPU. In this system so many alternatives still possible to map the association rule mining problem to GPU processors. Hence, in future it tremendous chances to improve the GPU processing using proper allocation of a problem using great memory utilization and less time-consuming

REFERENCES

[1] Han, J., Kamber, J. and Pei, M. (2011) Data Mining: Concepts and Techniques, Vol. 8, 3rd ed., pp.1–50, China Machine Press.

[2] Agrawal, R. and Shafer, J. (1996) Parallel mining of associations rules, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp.962–969.

[3] Bessedik, M., Bouakline, T. and Drias, H. (2011) How can bees color graphs, Int. J. Bio-Inspired Computation, Vol. 3, No. 1, pp.67–76.

[4] Pham, D.T., Castellani, M. (2009), The Bees Algorithm – Modelling Foraging Behaviour to Solve Continuous Optimisation Problems. Proc. ImechE, Part C, 223(12), 2919-2938.

[5] Drias, H., Sadeg, S. and Yahi, S. (2005) 'Cooperative bees swarm for solving the maximum weighted satisfiability problem,' in Proceedings of IWANN, pp.318–325.

[6] Agrawal R, I mielinski T and Swami A, Mining association rules between sets of items in large databases, Proceedings of the ACM SIG2 MOD,Washington DC, pp 207- 216, 1993.

[7] Agrawal, R. and Ramakrishnan,S.: Fast algorithms for association rules in large databases (http://rakesh.agrawalfamily.com/papers/vldb94apriori.pdf), in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Cqarlo; editors, Proc of the 20th International Conference on very large Data bases -VLDB), Santiago, Chile, PP 487-499, Sept 2004.

[8] Han, J., Pei, J., Yin, J., Mai, R.: Mining frequent patterns without candidate generation, in Data Knowledge and Knowledge discovery, No 8, PP 53-87, 2004.

[9] Agrawal R, Shafer JC (1996) Parallel mining of association rules. IEEE Trans Knowl Data Eng 8(6):962–969.

[10] Melab N, Talbi E-G (2001) A parallel genetic algorithm for rule mining. In: Proceedings of the 15th international parallel and distributed processing symposium. IEEE Computer Society, London

[11] Cano A, Luna JM, Ventura S (2013) High-performance evaluation of evolutionary-mined association rules on GPUs. J Supercomput 1–24

[12] Djenouri, Y., Drias, H., Habbas, Z., Mosteghanemi, H.: Bees Swarm Optimization for Web Association Rule Mining. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 3, pp. 142–146. IEEE (2012)

[13] Djenouri, Y., Drias, H., Chemchem, A.: A hybrid Bees Swarm Optimization and Tabu Search algorithm for Association rule mining. In: 2013 World Congress on Nature and Biologically Inspired Computing (NaBIC). IEEE (2013)

[14] Djenouri, Y., Bendjoudi A, Mehdi M, Nadia N, Habbas Z : GPU-based bees swarm optimization for association rule mining in: J Supercomput(2015) 71:1318-1344

[15] K. Bhuvaneswari and K. Saravanan, Privacy Preserving Association Rule Mining From Highly Secured Outsourced Databases. International Journal of Computer Engineering & Technology, 8(4), 2017, pp. 98–107.

[16] Rahul Shajan and Gladston Raj S, Association Rule Mining Based Analysis on Horoscope Data – A Perspective Study. International Journal of Computer Engineering & Technology, 8(2), 2017, pp.76–81.

[17] Aruna J. Chamatkar, A Study on Association Rule Mining with Neural Based Framework, International Journal of Computer Engineering & Technology (IJCET), Volume 5, Issue 9, September (2014), pp. 172-181

[18] Rajesh V. Argiddi, A Study of Association Rule Mining in Fragmented Item-Sets for Prediction of Transactions Outcome in Stock Trading Systems, International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 2, July- September (2012), pp. 478-486