ANALYSING THE EFFICASY OF GWO AND PSO IN DEVELOPING OPTIMISED MODEL OF C5.O LINKED TO ASSCOCIATION RULES IN DETERMINING OPTIMISERS TO PREDICT EMPLOY ATTRITION

MEHUL SHOREWALA
Delhi Public School, Mathura Road

ABSTRACT

Predicting the attrition of employee based on 5 selected attributes which are Gender, Distance from Home, Environment Satisfaction, Work Life Balance and Education Field out of 36 variables present in the dataset. Application of Grey Wolf Optimisation (GWO) Algorithm and Particle Swarm Optimisation (PSO) on the model of Decision Tree Algorithm "C5.0" which is fed in the inputs of Associated Rules, using this optimized algorithm for the prediction of employee attrition using IBM Watson Human Resource Employee Attrition Data. After comparing the efficiency of GWO and PSO, we have come to a conclusion that time to predict an employee attrition and consumption of RAM have been optimized with GWO. Employee Attrition is one of the major problems faced by companies now-a-days. Sometimes, when the long term working employees leave the company, it affects the relationship of the company with the client and in turn affects the revenue of the company if the person replacing the old employee isn't able manage a good rapport with the client. The paper can be used to frame better work policies which will help both the employer and employee. It can be seen as a mirror to the working conditions of the employees.

Keywords: Apriori Algorithm · Association Technique · C5.0 · Data Mining · Decision Tree · Employee Attrition · Entropy · IBM Watson HR · Information gain · Grey Wolf Optimization · Particle Swarm Optimization

1. INTRODUCTION

Employee Attrition is one of the major problems faced by companies now-a-days. Loss of employees from a company is actually the loss of all the training and efforts put in by the company in the employee. Sometimes, when the long term working employees leave the company, it affects the relationship of the company with the client and in turn affects the revenue of the company if the person replacing the old employee isn't able manage a good rapport with the client. Also, finding an immediate replacement for the leaving employee is difficult and company has to put in the time and efforts in hiring new people leading to loss of valuable time and resources.

Determining the attrition rate helps a company compare it with the industry average and work towards reducing the attrition rate. It helps the company in knowing the reasons for attrition of employees so that they can improve as a company and keep its employees satisfied and content.

Here comes the role of Data Analytics that helps us use various factors like Work Life Balance, Environment Satisfaction and other factors to predict the attrition of an employee.

In our Research Work, we have used Data Mining techniques to predict the attrition of an employee using a model which is optimised with the help of Grey Wolf Algorithm.

2. LITERATURE REVIEW

In the year 2012, Sheila A.Abaya focused on the limitation of Apriori algorithm and provided an innovative method for the improvement of algorithm by introducing a set of rules to achieve the required outcome. In this algorithm number of candidate keys were minimised by using factors of set size and set frequency. The average results for execution time and the database pass improve by 38% and 33% respectively when using the modified algorithm.

In the year 2013, Leena Ragha and Jayshree Jha worked on data mining for educational purposes that is, data originating in the educational context. And for the purpose, Apriori algorithm was considered the most appropriate. On the basis of Apriori algorithm and its corresponding research and analysis, the referenced paper identifies the main issues associated with the algorithm’s application in data mining in educational context. and henceforth introduces a better performing algorithm based on support matrix, which uses the approach of bottom-up with standard deviation functional model for the purpose of mining data patterns occurring frequently in educational data.

In the year 2013, Jiao Yabing focused on the fact that whenever Apriori algorithm encountered high density data because of which a long patterns emerged in large numbers, the performance drastically reduced. So to fix this problem, improvements were proposed. In Apriori algorithm, C k-1, where C stands for candidate, is put in comparison towards support level after it is discovered. Item set that is less compared to support level is pruned. By doing this the connected item set number will be decreased and thereby the number of candidate items will decrease as well. This improved Apriori algorithm and its implementation were discussed in this paper.

In the year 2015, Rutvija Pandya and Jayati Pandya compared the algorithms ID3, C5.0and C4.5 with one another and it was found that the C5.0 algorithm gave much more efficient and accurate results. Hence in the light of that, C5.0 algorithm was used as base classifier. Therefore, it classifies with low memory use because it generates fewer rules. Further, due to low error rates, the accuracy is high as well.

Further, the C5 algorithm also performs feature selection which is a technique that works on the assumption that data contains many redundant features which provide no useful information and hence should be removed. And therefore only relevant features that are beneficial to model construction should be selected. In addition, the reduced error pruning also solves the decision tree's problem of over fitting.

In the year 2015, Seyedali Mirjalili with other research scholars proposed a new technique inspired by grey wolves called Grey Wolf Optimizer hereby shortened to GWO. This algorithm is inspired by the hunting methodology of grey wolves and their way of leadership in nature. For the purpose of simulation alpha, beta, omega, and delta type of grey wolves are taken. Also the three steps, that is, searching, encircling and attacking the prey, utilised for the purpose of hunting by wolves are implemented in the algorithm. After that, benchmarking was done against 29 test functions. In addition, comparison was done against differential evolution, particle swarm optimisation, evolution strategy and other such algorithms. The results of comparison were found to be very good. Further, in order to show that the proposed algorithm works well in challenging issues with many unknowns, the paper attempts its application in a variety of classical design and engineering problems.

In the year 2015, Dr. Sudhir Sharma with other research scholars presented a new way to provide a solution for economic load-dispatch problem (convex). This was done by using a grey wolf inspired meta heuristic known as grey wolf optimization. The purpose of economic load dispatch is to reduce as far as possible the generation cost even as it satisfies the multiple constraints it is subjected to, in the process of supplying the required loads requested by power systems. The above said technique was applied on different systems for testing, with varying load demands. To demonstrate the improvement when using GWO, the results were compared with existing systems and techniques and the findings showed considerable improvement in performance parameters when using Grey Wolf Optimization. Not just that, the Grey Wolf Optimization was also found to be reliable, simple and efficient in nature.

In the year 2011, Dian Palupi Rini, Siti Mariyam Shamsuddin, Siti Sophiyati Yuhaniz studied about Particle swarm optimisation or PSO in short is an optimisation and computational search method which draws its inspiration from biology. It's development took place in 1995. It was developed by Eberhart and Kennedy. The social behaviours of fish schooling and birds flocking formed its basis. The algorithm draws inspiration from from behaviour of those animals who do not have any leader in their swarm or group, such as fish schooling or bird flocking. The reason behind that lies in the fact that such flocks with no leaders, find food at random and then will follow an animal of the group that is nearest to the source of food ie a solution (potential). The flocks will communicate amongst themselves about who has a better solution already. And hence the member that has the better solution will provide that information to the rest of the flock as well and hence others will also come to that place. This process will be repeated untill the best solution is discovered. This is the process that particle swarm optimisation follows.

3. MATERIALS AND METHODS

We have analysed IBM Watson Human Resource Employee Attrition Data (source – Kaggle) set to predict the employee attrition based on 5 selected bases classes, which are Gender, Distance from Home, Environment Satisfaction, Work Life Balance and Education Field variables out of the set of 36 variables.

Tools used are Microsoft Visual Studio and Microsoft SQL Server on core i7, 7th Generation processor with 16GB RAM.

3.1 Association Technique

In 1993, Agrawal introduced association rule to track down relationship between products that belong to a set of transactions recorded in a supermarket. Since a large number of rules can be determined, there is a need to track down rules that are relevant. Constraints such as support and confidence are used to filter significant rules. By support, we mean probability that an item set P is present in the set of transactions T and confidence refers to the proportion of transactions that contain data item Q which contains data item P with respect to all transactions containing data item P.

3.1.1 Apriori Algorithm

Apriori algorithm counts item-sets by using breadth-first search algorithm along with Hash-tree structure. It uses a bottom-up approach, extending each item set as it moves upward. First, item-sets of length k are generated from a set of k-1 item-sets. After that, infrequent sub-patterns are pruned. Further, transactional database is scanned to determine frequently occurring item-sets among the extended item-sets. Apriori algorithm terminates when no further extension is possible.

3.2 Decision Tree

A Decision Tree is used to make computations based on past data, these computations are used to make important decisions taken by executives in an industry.

3.2.1 C5.0

It is used to analyse data sets that contain thousands of entries, C5.0 algorithm is represented in the form of decision tree in order to make it more interpretable. C5.0 algorithm is better as compared to C4.5 in measures of speed, memory utilization and the size of decision tree.

3.2.2 Entropy

Information Entropy is defined as the amount of information achieved on the basis of probabilistic outcomes of data values which means if the probability of occurrence of a data value is less then it will render more information as compared to a data value having higher probability of occurrence. For each data value, information entropy associated is the negative logarithm of the probability mass function. Information entropy can be cited as the uncertainty which is usually measured in bits. For example, consider a set of events having different probability of occurrences. The event having lowest probability of occurrence will deliver more information as compared to other events.

3.2.3 Information Gain

It is defined as the variation in one probability distribution as compared to the other probability distribution. Kullback - Leibler divergence is the other name for Information Gain. Although information gain is a good measure to decide relevance of an attribute but it holds problem when it is applied to attribute that holds a large set of discrete values.

3.3 Grey Wolf Optimisation

Grey wolf optimization which was originated by Mirhalili et al., 2014 that represents how wolves perform hunting based on their hierarchal leadership. Grey wolf always live in a pack where their leader can be a male or female often known as Alpha, who is mainly responsible for decision making. Alpha is followed by Beta, Delta and Omega in a hierarchal manner. All of these groups help their senior members for decision making. This hunting technique and hierarchy of wolves are mathematically modelled to develop Grey Wolf Optimization technique. GWO has a few parameters and is easy to implement as compared to other swarm optimisation techniques.

3.4 Particle Swarm Optimisation

In 1995, Particle Swarm Optimization (PSO) has been designed by Russell Eberhart and James Kennedy which is enlivened by the flocking examples of fish and birds. Initially, these two began creating PC programming reproductions of winged creatures flocking around nourishment sources, then later acknowledged how well their calculations took a shot at optimisation issues.

Particle Swarm Optimization may sound complicated, however it's an exceptionally simple algorithm. Over various cycles, a gathering of factors have their values balanced nearer to the part whose value is nearest to the objective at any given time.

Envision a rush of feathered creatures hovering over a region where they can notice the smell hidden source of food. The one who is nearest to the nourishment twitters the loudest and other birds swing around towards it. In the event that any of the other revolving around birds comes nearer to the objective than the principal, it tweets louder and the others veer over toward him. This fixing design proceeds until the point when one of the flying creatures chances upon the nourishment. It's a calculation that is basic and simple to actualize.

4. PROPOSED RESEARCH SCHEME

The dataset used has been acquired from Kaggle. After acquiring the dataset, we selected base classes for the employee attrition which are: Gender, Distance from Home, Environment Satisfaction, Work Life Balance and Education Field. Then we have cleaned and transformed the dataset according to the selected attributes. During cleaning, we have removed entries that contain redundant values and incomplete tuples.

Then to use our new approach that is C5.0 with association, we have applied association rule mining using Apriori algorithm to form association rules using selected attributes. Then, using these association rules we have trained the C5.0 decision tree. Using this model, we have then predicted the attrition of employee and then matched the predicted results with the actual attrition to evaluate the efficiency of the proposed algorithm on this dataset.

Now, to further optimise our model, we have used Grey Wolf Optimiser. Then using this optimised algorithm, we have again predicted the attrition of employee and then matched the predicted results with the actual attrition to evaluate the efficiency of the optimised model on this dataset.

To compare Grey Wolf Optimiser with another optimisation algorithm, we have also used Particle Swarm Optimisation and again predicted the attrition of employee.

5. PERFORMANCE COMPARISON

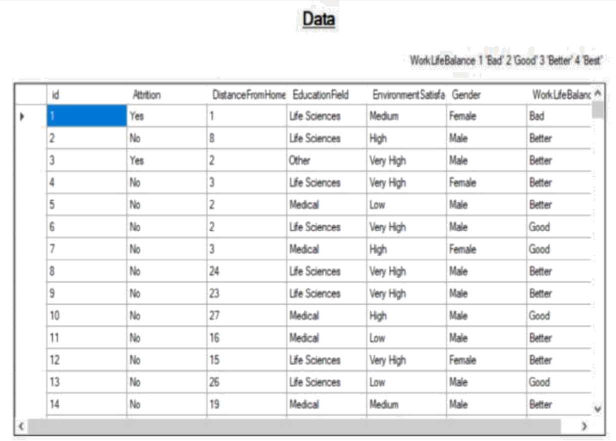
After acquiring the dataset from Kaggle, we are using GWO and PSO to optimize further C5.0 with association algorithm. We have observed that the time and memory consumption is least with GWO, then with PSO and then finally with Association Rules in comparison to traditional C5.0. Therefore, Grey Wolf Optimised C5.0 with association algorithm is more efficient in time and memory consumption as compared to others techniques with C5.0 as shown in the table 1.1.

Table 1.1 It shows improved efficiency when Grey Wolf Optimised C5.0 Association Algorithm is used as compared to the PSO Optimised and C5.0 with Association Rule Algorithm

Basis	GWO Optimised	PSO Optimised	C5.0 with Association Rules	Traditional C5.0
Processing Time consumption	0.046ms	0.14ms	0.2ms	2ms
RAM consumption	30.9MB	32.5MB	39.6MB	48MB

6. RESULTS AND DISCUSSION

After comparing the efficiency of different algorithms, we have come to a conclusion that time to predict an employee attrition and consumption of RAM have been optimized* with GWO. Figure 6.1 shows the transformed data set after redundancies are removed. Figure 6.2, 6.3 shows Entropy and Attrition values while training and testing respectively. In figure 6.5 shows actual and predicted values of Attrition. Figure 6.4 and 6.6 shows correctly and incorrectly predicted values of employee Attrition. Figure 6.7 and 6.8 shows Processing time and RAM consumption by optimized Grey Wolf Algorithm.



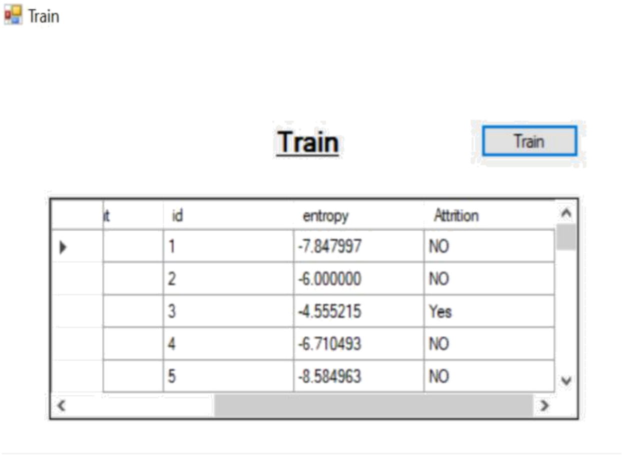
Data

WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

id	Attrition	DistanceFromHome	EducationField	EnvironmentSatisfa	Gender	WorkLifeBalanc
1	Yes	1	Life Sciences	Medium	Female	Bad
2	No	8	Life Sciences	High	Male	Better
3	Yes	2	Other	Very High	Male	Better
4	No	3	Life Sciences	Very High	Female	Better
5	No	2	Medical	Low	Male	Better
6	No	2	Life Sciences	Very High	Male	Good
7	No	3	Medical	High	Female	Good
8	No	24	Life Sciences	Very High	Male	Better
9	No	23	Life Sciences	Very High	Male	Better
10	No	27	Medical	High	Male	Good
11	No	16	Medical	Low	Male	Better
12	No	15	Life Sciences	Very High	Female	Better
13	No	26	Life Sciences	Low	Male	Good
14	No	19	Medical	Medium	Male	Better

Fig.6.1 Transformed Dataset

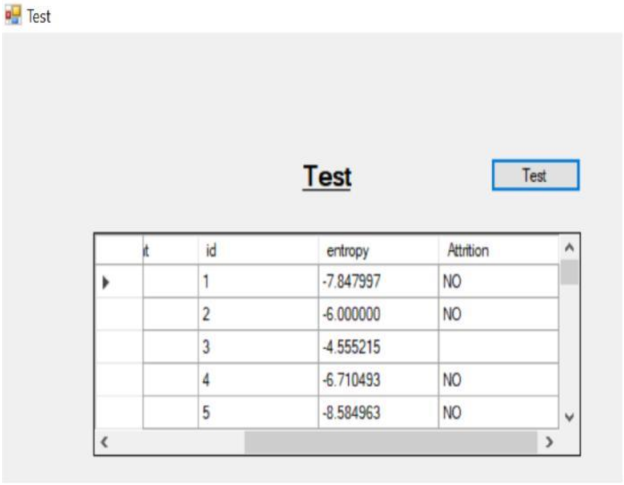
while training



Train

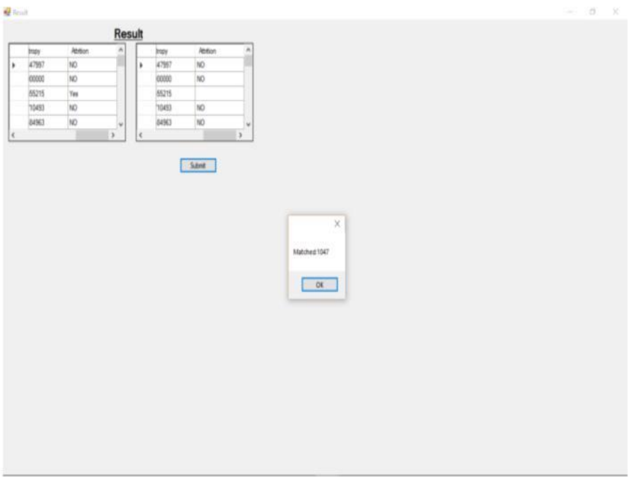
t	id	entropy	Attrition
	1	-7.847997	NO
	2	-6.000000	NO
	3	-4.555215	Yes
	4	-6.710493	NO
	5	-8.584963	NO

***Fig.6.2.** Entropy and Attrition values*



	t	id	entropy	Attrition
▶		1	-7.847997	NO
		2	-6.000000	NO
		3	-4.555215	
		4	-6.710493	NO
		5	-8.584963	NO

Fig. 6.3 *Computation of Entropy and attrition values while testing*

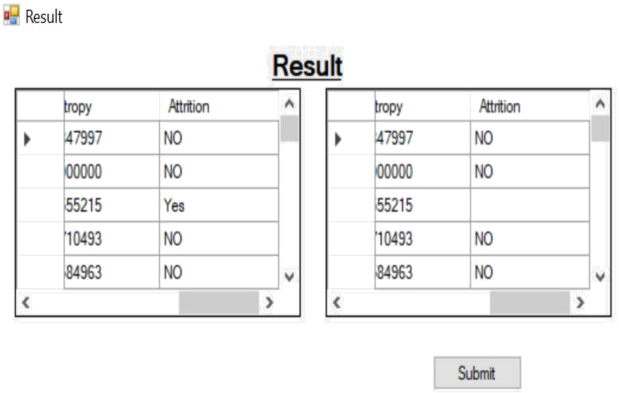


	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	Yes
	10493	NO
	84963	NO

	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	
	10493	NO
	84963	NO

Matched:1047

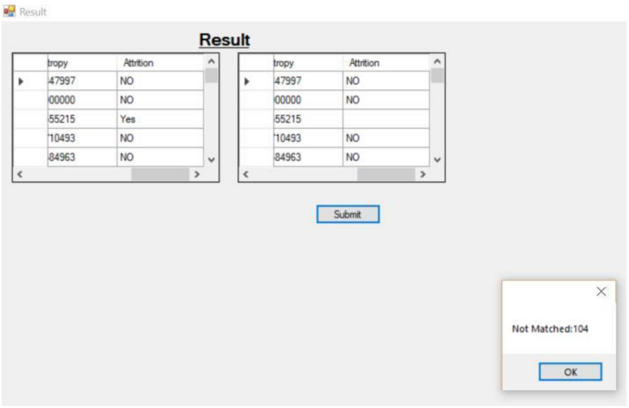
Figure 6.4 *Correctly predicted instances by the Grey Wolf optimised Algorithm*



	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	Yes
	'10493	NO
	84963	NO

	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	
	'10493	NO
	84963	NO

Fig. 6.5 *Result of prediction made by Grey Wolf optimized Algorithm. The Grid View on the left of the Image shows the actual values of attrition while the Grid View on the right shows the predicted values of attrition. Blank cell in the column shows that the prediction was incorrect*

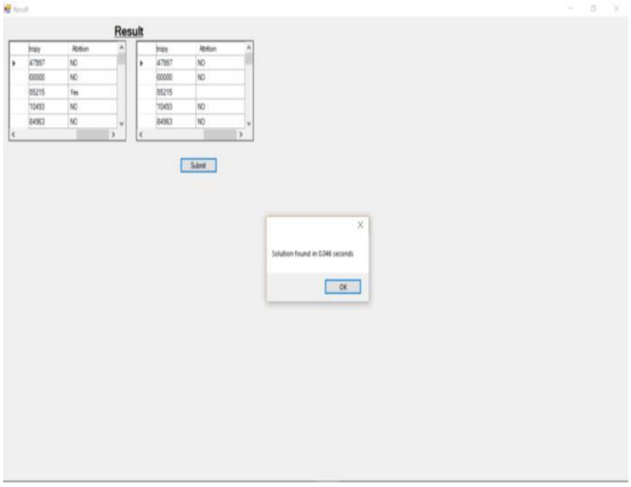


	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	Yes
	'10493	NO
	84963	NO

	tropy	Attrition
▶	47997	NO
	00000	NO
	55215	
	'10493	NO
	84963	NO

Not Matched:104

Fig. 6.6 *Incorrectly predicted instances by the GWO Algorithm*



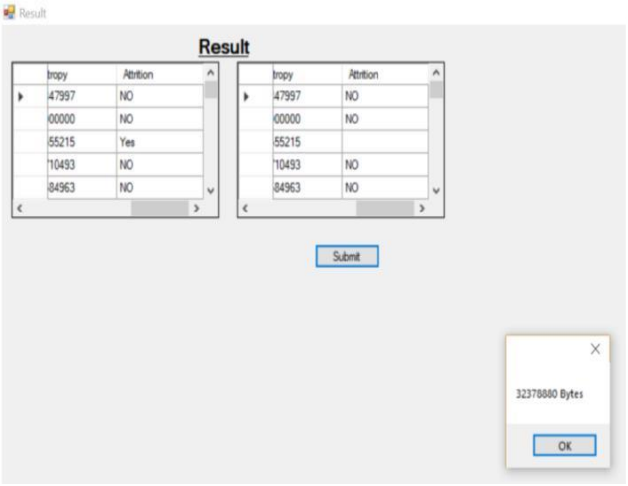
Result

	tropy	Attrition
▸	47997	NO
	00000	NO
	55215	Yes
	10493	NO
	84963	NO

	tropy	Attrition
▸	47997	NO
	00000	NO
	55215	
	10493	NO
	84963	NO

Solution found in 0.046 seconds

Fig. 6.7 *Processing time consumption by the Grey Wolf optimised Algorithm*



Result

	tropy	Attrition
▸	47997	NO
	00000	NO
	55215	Yes
	10493	NO
	84963	NO

	tropy	Attrition
▸	47997	NO
	00000	NO
	55215	
	10493	NO
	84963	NO

32378880 Bytes

Fig. 6.8 *RAM consumption by Grey wolf optimization algorithm*

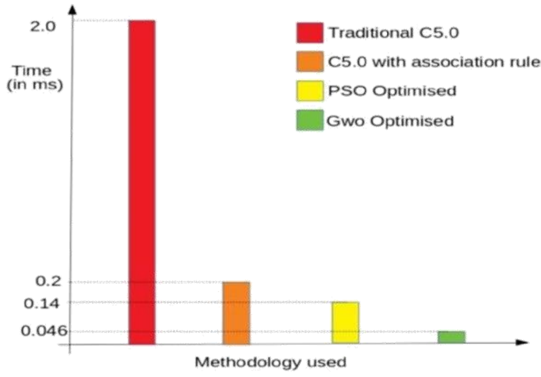


Fig. 6.9 *Process Time Consumption*

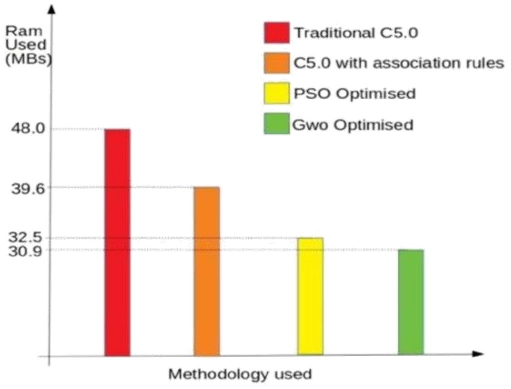


Fig 6.10 *Ram Consumption*

7 FUTURE WORK

The work done in this paper holds immense scope for extension in future endeavours. This is because a number of nature inspired algorithms are currently active areas of research. Furthermore, the grey wolf optimizer algorithm utilised in this paper itself has the potential to be superseded by another nature based algorithm.

Nature based algorithms is a very prolific research area. This is because the problems with which we are normally familiar are getting further complicated and complex due to size and other aspects. In addition, new problems are cropping up in which existing methods are not effective. Nature seems have faced similar problems and solved them in due course of time. That is the reason we get a lot of inspiration from it . Some of the recent nature inspired algorithms are Artificial bee colony algorithm, The firefly algorithm, the social spider algorithm, the bat algorithm, the strawberry algorithm, the plant propagation algorithm, and so on. These are very effective compared to early nature inspired algorithms such as the genetic algorithm, ant colony and swarm optimization and so on. Such algorithms have very few parameters that need arbitrary setting.

8 CONCLUSION

After comparing the efficiency of Grey Wolf Optimised C5.0 with association algorithm to C5.0 with association rule and Particle Swarm optimization, it is observed that the time to predict an employee attrition and consumption of RAM is optimized hence it can be concluded that efficiency of algorithm is improved when C5.0 Association rule algorithm is used with Grey Wolf.