DOI: 10.21276/sjet

Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2017; 5(7):335-339 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

An Implementation of Sequential Rule Mining Using Mapreduce Based Genetic Algorithm

Amanjeet Kour[1], Om prakash Dewangan[2], Toran verma[3]
[1,2,3]Department of computer Science, Rungta College Of technology, Bhilai (C.G), India

***Corresponding author**
Dr. Amanjeet Kour
Email: amanjeet.k.b@gmail.com

Abstract: Sequential rule mining is a fundamental technique of data mining which has many application one of which is in the area of bioinformatics. Bioinformatics is an application of information technology to store gigantic biological data in an organised way so that the data can easily be analysed. These biological data is in form of sequences of proteins and sequence analysis is one of the major research areas in bioinformatics. These Bioinformatics data are incremental datasets and at the present era of cheap information technology these bioinformatics data has become big data. Efficient mining of big data require parallel as well as iterative techniques. Here we propose a technique to analyse these sequential data and extract sequential rule using map reduce Framework of Hadoop mounted over genetic algorithm. Map reduce is responsible for parallelizing of mining technique where as genetic algorithm would work in an iterative manner to generate sequential rules from DNA sequences.
Keywords: Data mining, rule mining, Big data, Hadoop, Mapreduce, Genetic Algorithm.

INTRODUCTION

Sequential rule mining is a technique of data mining applicable only to sequential database. A sequential database contains list of sequential data. Following is small set of sequences named as seq1, seq2, seq3, seq4 containing symbols “a”, ”b”, ”c”, ”d”, ”e”, ”f”, ”g”, ”h”. Sequence is a well-ordered list of elements.

Table: sequential database containing four sequences

ID	Sequences
Seq1	<{a,b},{c},{f},{g}.{e}>
Seq2	<{a,d},{c},{b},{a,b,e,f}>
Seq3	<{a},{b},{f},{e}>
Seq4	<{b},{f,g,h}>

Sequences are very ordinary type of data structure seen in many different domains such as bioinformatics (DNA sequences), clicks on websites, sentences in text. Commonly finding of sequential pattern is done through various techniques of data mining but sequential rule mining is rarely done. Sequential pattern is a subsequence that appears in several sequences of database selected on the basis of support of the sequence. A more appropriate updating of sequential pattern mining was sequential rule mining which is finding of a rule of the form X->Y where X and Y are set of items. We find such rules by using measures of support and confidence of that particular rule.

- Support: number of sequence where X appears before Y, divided by number of sequences.
- Confidence: number of sequences where X happens before Y, divided by number of sequences where X occurs.

Bioinformatics is a field of information system where sequential data is largely generated. Here biological information is found in sequential pattern such as sequence of protein but the data found in bioinformatics is very large in size. As the bioinformatics data is voluminous as well as incremental it is also very complex. These data can never be analysed without the help of mining techniques. The normal mining techniques could help only to an extend and would fail in its efficiency after a limit. A parallelized as well as iterative way is required to solve this big data problem.

Map reduce a framework of Hadoop is supposed to be most efficient way of parallelizing any task. Hadoop is a frame of tools particularly designed with an intention to deal with the issue of big data. Hadoop is an open source collection of tools distributed under the license of apache i.e. no particular company holds Hadoop and is maintained by apache. The key logic behind the development of Hadoop is the handling of big data. Big data has 3 V's of challenges; they are velocity, volume, Variety where velocity refers to great increase in data every day resulting in large volume of data and variety explains that we have unorganised and diverse type of data. Traditional enterprise approach used single powerful computer to solve the big data issue which had a limit and could not show efficiency after that limit. So Hadoop brought up a veryunique approach i.e.map reduce. This mapreduce is capable of breaking down the data into smaller blocks and then deal with it. Similarly it breaks the computations to smaller blocks and sends these small blocks of tasks or data to various nodes to be used for processing and later on all of them are collected and united to form the solution.

To do the above work Hadoop has an efficient framework -mapreduce. Here two functions work one after the another one is mapper and another is reduce. Mapper is applied to input data were as reducer is applied to intermediate results that come from mapper. Hadoop shuffles and regroups the data according to key value pair. The prerequisite for mapreduce is that all the data must be arranged in key value pair <k1, v1> which becomes the basic unit of the framework. Hadoop replies applies the map() to where ever data is sitting in form of key value pair as<k1,v1> and generation of list of<key, value> pair takes place through map() referred as <list(k2,v2)>. Hadoop takes output of map () and will shuffle it, group it according to the key as <K2, list(v2)> After this the Hadoop would proceed the intermediate results to reducer() giving out the proper result as <list(k3,v3)>.

Genetic algorithm is an approach which mimics the biological evolution to solve engineering problem. First thing that is done in this algorithm is we create a set of random population. This population is the set of all the possible solutions of a problem. These population or solutions are considered as parents and they undergo combination and mutation to produce children population. For each candidate of the population some fitness value is calculated. This fitness value represents how fit is the solution to the problem. This approach further works on the basis of Darwin's theory i.e. "survival of fittest". The fitter are selected to be in the population and unfit solutions are dropped and the fitter go for further combination and mutation to generate offspring. In this way group of optimal solutions are generated using the genetic algorithm.

LITERATURE REVIEW

When discussion rises about the topic of big data the popular topic of Information industry, so many challenges and also solution of problems to some extend is seen. MR Pre Post [1] method is found for mining big data which is a hybrid method that has imposed map reduce over the Pre Post technique of mining and proved that the idea is good for mining large data. In realworld the problem of big data analysis for heart disease detection was taken care of in [4] through map reduce technique. Here the data from a centralized medical data system is taken and is analysed. In the analysis process particular patient record in categorized into of the classification among the two. The Two classifications are normal and abnormal. This detection step is done through map reduce technique by detecting the disease and reducing the dataset.

If we move on our discussion from big data mining to mining of sequential data for finding sequential rule then we saw Rule Growth and CM Rules [2,3] algorithm for mining sequential rules. Both the techniques work under the category of mini g several sequences. Rule growth works over pattern growth approach forextracting rules efficiently were as CM Rules method is based on association rule mininghence can discover both association rule as well as sequential rule. In [2] Rule Growth was compared to CM Rules and showed that Rule Growth clearly outperformed CM Rules.

We also found a concept of closed sequential pattern which was generated using genetic algorithm. The proposed algorithm is named s G-CSPM [5]. Although previously few more techniques were also introduced to generate closed sequential patterns such as Clo Span and BIDE, G-CSPM had less time complexity than others. This was the first method to utilize genetic algorithm for closed sequential mining and also worked over fitness function and pruning method. In the result of [5] it was clearly shown that G-CSPM outperformed Clo Span.

Genetic algorithm that we discussed about in the above lines is an optimization technique and was used in [6] optimization of association rule mining process. Here first of all strong association rule are generated by combining two techniques Apriori algorithm and FP growth algorithm. Later on genetic algorithm is applied to optimize the result .Genetic algorithm helps in finding the rule of high interestingness.

IN the above review of literature following things are seen

- Big data problem can be very efficiently solved by applying map reduce technique.

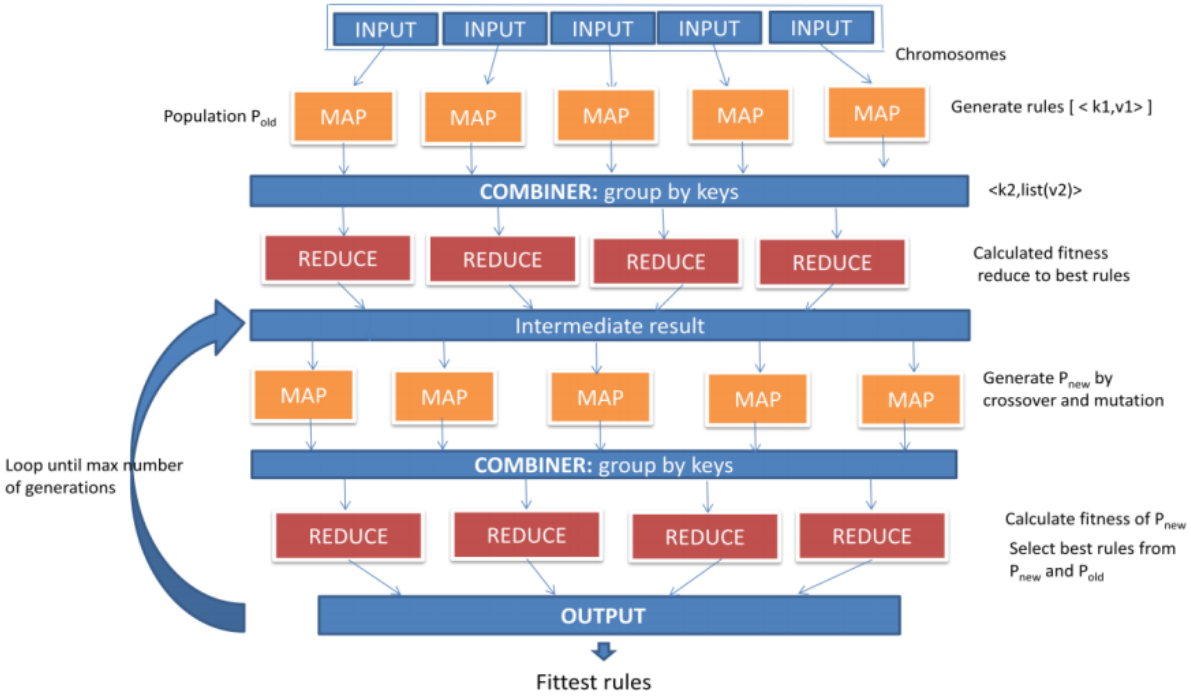
- Genetic algorithm can be implemented for two uses may be for optimization as in [6] or for directly generating the rules as in [5].

In our proposed work we have made an attempt to hybridize genetic algorithm with map reduce to solve the big data problem. Genetic algorithm will be used to perform the duty of generating rules rather than only optimization work.

METHODOLOGY

The proposed model of parallelization of genetic algorithm through map reduces technique of Hadoop framework is best described through the flow diagram below. The proposed approach is divided into two rounds of map reduce. First round is responsible for generating chromosomes from the input data. These chromosomes will help to provide initial population i.e. all the possible solutions. These solutions are the possible rules that can be generated from the sequential data. These rules are mapped and reduced in the first round to select the fittest of all. Selection is done on the basis of fitness values of each rule provided a minimum fitness value.

Round two of the framework works over the fittest set of rules generated in the first round. Output of the first round is input to the second round. These set of rules are considered as parent population and undergo cross over and mutation to produce next generation of population. Again fitness each child solution is also calculated. At last the fittest from the parent as well as children are selected for future generation. The second round is repeated until maximum number of generations is reached. Hence we receive the fittest or best rules from the sequential data



Round 1:

Mapper: First step of round one where input data is provided as input to Mapper. Here each line is read and chromosome of the genetic algorithm are generated. Set of all chromosomes are the population P old. From each chromosomes the rules are generates. The population is generated by random initialization.

Map: input→<k1,v1>

Combiner: Combiner received the output from the mappers and groups all the rules according to key value grouping concept of map reduce technique of Hadoop. Values of all the keys are summed up and the key value sets are forwarded to reducer. It is also called 'local reducer' as it reduces the task of reducer by minimizing data exchange.

Combiner: <k1,v1>→<k2,list(v2)>

Reducer: Reducer sums up the list(v2) generating <k3,v3> set. Additional work that the reducer does in this framework is that using this value v3 of each key it calculates the fitness of each key. The key holds the rules that are solution to our problem. The fitness is calculated by fitness function.

Reducer:<K2,list(v2)>→<k3,v3>

Fitness function: The fitness of each key is calculated on the basis of value v3. As we talk about finding of fittest sequential rule, support is the measure that denotes the fitness of the sequential rule.

Sup= val /N

Val: V3 the summed up value generated in the reducer itself.

N: N is the population size which we have defines to be 22. after the fitness is calculated for each rule, the rules with higher fitness are selected where minimum fitness criteria is set. The rules having fitness less that minimum fitness are rejected from the pool of solutions and rest best rules are the intermediate result.

Min_fitness=0.020

Round Two

Mapper: The best rules are again mapped. These rules are the parents of the first generation which is used to generate next generation of population through cross over and mutation. Single point cross over and bit flip mutation is performed for finding the new generation named as P new. Again the key and values are set for this P new.

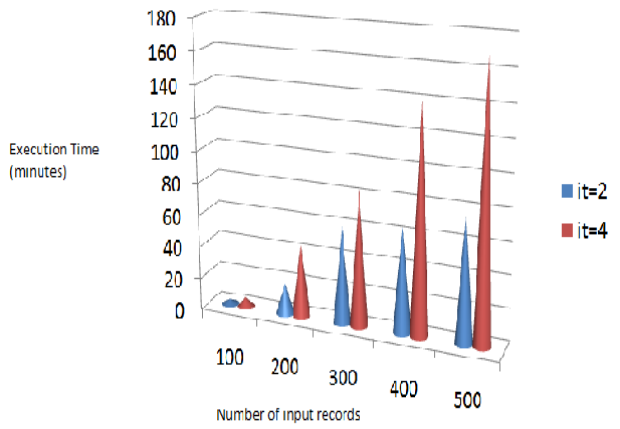
Combiner: Combiner receives the output from mapper and groups the population by key and list of values corresponding to each key is given as output by the combiner.

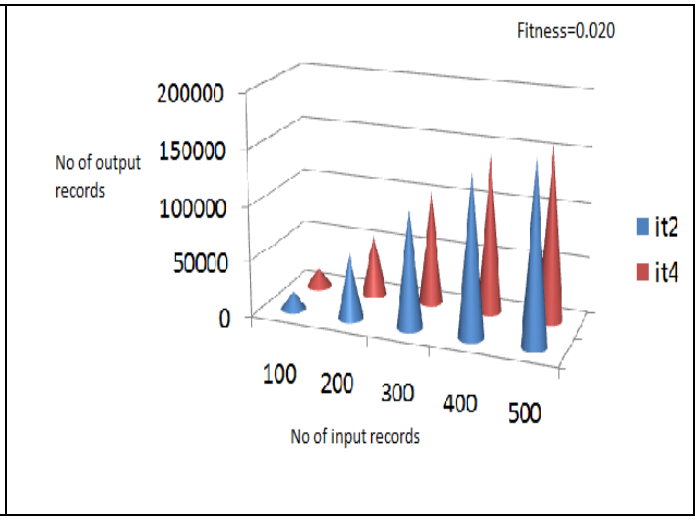
Reducer: Reducer again calculates the fitness value for P new as it was done for parent generation. All the rules having fitness greater than min-fitness are generated and best rules are selected from the pool of P old and P new .selected fittest from both P old and P new become the new generation .The output from the reducer of the second round is proceeded to intermediate result.

Round two is repeated until maximum number of generation is reached and thus final set of best rules are generated.

RESULT

Four experimenting our framework the data is taken from GIL from Indiana university of Biology from evaluation of gene structure prediction program, genomics,34:353-357(1996) [7]. Following are the graphs representing the output of our work.





As we can see in the first graph, plotting is done between number of input records and the time taken for execution. Here it represents the number of iteration i.e. number of generations created by cross over and mutation. We have plotting for two situations one is when only two generations are used to create the result and other when four generation used to create the result. We observed that double the iteration double the time it takes to execute the work. Second graph is plotted between number of input and the number of output generated as rules. For both the situations, minimum fitness is kept fixed i.e. 0.020 for both iteration two and iteration four, result generated is same but time it took for execution is just double. From the result we would conclude that the system gives same appropriate result in less time and less iteration that it would give in more time and iteration.

CONCLUSION

By using the map reduce framework of Hadoop we have parallelized the big data processing required in bioinformatics. The work done is DNA sequence analysis which results in sequential rule mining. The rules generated can be of great importance to medical

science such as determining diseases, symptoms and their treatment. The rule mining is performed when genetic algorithm worked parallely through the use of map reduce to analyse large data. Map reduce first splits the data into small blocks and distributes it to multiple nodes. At each node the blocks are analysed using genetic algorithm through various iterations hence genetic algorithm hybridized with map reduce is responsible for rule generation followed by rule optimization. In future this frame work can be used in various other areas to generate result from big data.

REFERENCES

1. Liao J, Zhao Y, Long S. MRPrePost—A parallel algorithm adapted for mining big data. InElectronics, Computer and Applications, 2014 IEEE Workshop on 2014 May 8 (pp. 564-568). IEEE.
2. Fournier-Viger P, Nkambou R, Tseng VS. RuleGrowth: mining sequential rules common to several sequences by pattern-growth. InProceedings of the 2011 ACM symposium on applied computing 2011 Mar 21 (pp. 956-961). ACM.
3. Fournier-Viger P, Faghihi U, Nkambou R, Nguifo EM. CMRules: Mining sequential rules common to several sequences. Knowledge-Based Systems. 2012 Feb 29;25(1):63-76.
4. Vaishali G, Kalaivani V. Big data analysis for heart disease detection system using map reduce technique. InComputing Technologies and Intelligent Data Engineering (ICCTIDE), International Conference on 2016 Jan 7 (pp. 1-6). IEEE.
5. Raju VP, Varma GS. Mining closed sequential patterns using genetic algorithm. InAdvanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on 2014 May 8 (pp. 634-637). IEEE.
6. Patel UK. Optimization of Association Rule Mining Using Genetic Algorithm. Optimization. 2016 Jun.
7. http://genome.crg.es/datasets/genomics96/#SEQS