Middle-East Journal of Scientific Research 19 (6): 807-810, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.19.6.1473

Mining Frequent Itemsets Using Genetic Algorithm

D. Kerana Hanirex and K.P. Kaliyamurthie

Bharath University, Chennai, India

Abstract: Association rule plays a vital role for mining frequent itemsets which is the recent research area in data mining. Frequent itemsets are generated by using partition, sampling and hashing technique. By using Genetic Algorithm (GA) we can improve the efficieny of finding frequent itemsets. The advantage of Genetic Algorithm is to find the frequent itemsets with less time complexity. The main aim of this paper is to find the frequent itemsets from the Dengue Virus data sets. Our findings reveals that Leucine (L), Phenylalanine (F), Lysine (K), Serine (S) and Glycine (G) are the dominating amino acids in Dengue Virus Type-1.genetic algorithm.This paper reveals the same dominating amino acids with less time in seconds.

Key words: Genetic Algorithm (GA) · Association Rule · Frequent itemset · Support · Confidence · Data Mining

INTRODUCTION

Mining association rule is one of the recent data mining research. Mining useful information and helpful knowledge from the large databases leads to an important research area [1, 2]. The Apriori algorithm is the basic algorithm for mining association rules.Data mining handles large amounts of data. The information can be used for applications such as market analysis, medical diagnosis and fraud detection. Given a set of items $I=\{I_1,I_2,\ldots In\}$ and a database transactions $D=\{t_1,t_2,\ldots\ldots t_m\}$ where $t_i=\{I_{i1},I_{i2}\ldots I_{in}\}$ and Iij $\in$ I, an association rule is of the form $X \Rightarrow Y$ where X, Y⊂I are sets of items called itemsets and $X \cap Y=\Phi$. The confidence or strength (α) for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain X∪Y to the number of transactions that contain X. The association rule problem is to identify all association rules with a minimum support and confidence. The most common approach to find association rules is to break up the problem into 2 parts.

- Find Large Itemsets
- Generate rule from the frequent Itemsets

A Large (Frequent) Itemset is an Itemset whose number of occurance is above the threshold (s) [3], the Apriori Algorithm is the most well known association rule algorithm and it is used in most commercial products. It uses largest itemset property [1].

"Any subset of a large itemset must be large" The basic idea of Apriori algorithm is to generate item sets of a particular size and then scan the database to count these to see if they are large. L_i is used to generate next C_{i+1}. L represent Large Itemset. C represents candidate items. All singleton itemsets are used as candidates in the first pass. The set of large item sets of the previous pass, L_{i-1} is joined with itself to determine the candidates.

Steps Involved in Apriori Algorithm

Input: Input sequence from polyprotein datasets [Dengue virus 1] from Gen Bank: AAB 27904.1

Output: Set of frequent patterns

Method: Step 1:

- Find all frequent itemsets
- Get frequent items whose occurrence greater than or equal to the minimum support
- Generate candidate itemsets and prune the results

Corresponding Author: K.P. Kaliyamurthie, Bharath University,Chennai, India.

Step 2: Generate association rules which satisfy min.support and min.confidence

Genetic Algorithm: Genetic Algorithms (GAs) are heuristic search algorithm based on the ideas of natural selection and genetic. GA is an approach to inductive learning. GA works in an iterative manner. It uses fitness measure to solve the problem [4]. Standard GA uses genetic operators such selection, crossover and mutation. GA runs to generate solutions for successive generations. Hence the quality of the solutions in successive generations improves [5], the process is terminated when an optimum solution is found. The functions of genetic operators are as follows:

Selection: Use a fitness function to evaluate the current solution. Where fitness is a comparable measure of how efficiently a chromosome solves the problem at hand.

Crossover: Crossover develops new elements for the population by combining parts of two elements currently in the population.

Mutation: Alters the new solutions in the search for better solutions.

In Genetic algorithm the operators are repeatedly applied to a population. This generational process is repeated until a termination condition has been reached.

Common Terminating Conditions Are:

- A solution is found that satisfies minimum criteria
- Fixed number of generations reached
- Allocated budget (computation time/money) reached
- Manual inspection
- Combinations of the above

Simple Generational Genetic Algorithm:

- Choose the initial population
- Evaluate the fitness of each
- Repeat on this generation until termination
- Select the best individuals for reproduction
- Breed new individuals through crossover and mutation
- Evaluate the individual fitness of new individuals
- Replace least-fit population with new individuals

The paper [6] proposes three different measures for performance evaluation of GAs such as the *likelihood of optimality*, the *average fitnessvalue* and the *likelihood of evolution leap*. Therefore it follows from the above theoretical discussion that GA based solution provides significant improvement in computational complexity in comparison with Apriori algorithm and all its variants [7].

We first load the sample of records from the transaction database that fits into memory. An initial population is created consisting of randomly generated transactions [8], each transaction can be represented by a string of bits. Our proposed genetic algorithm based method for finding frequent itemsets repeatedly transforms the population by executing the following steps such as:

- Fitness Evaluation
- Selection
- Recombination
- Replacement

RESULTS

This genetic algorithm is implemented by taking the sample sequential datasets for Dengue virus 1(DEN1): Gen Bank: AAB2 7904.1 which consists of 777 amino acids.

Table 1: Comparison between Apriori algorithm and Genetic Algorithm

Confidence	Time taken in secs (Apriori algorithm)	Time taken in secs (GA algorithm)
90	.0020	.0021
80	.0012	.0012
70	.0010	.0010
60	.0008	.0008
50	.0010	.0010

The accuracy of the system is measured by the time it takes to find the association rule. The following table shows the time taken in seconds for different confidence measures [9-20].

CONCLUSION

We have dealt with a challenging association rule mining problem of finding frequent itemsets the GA based method. The method, described here is very simple and efficient one.This method almost working with similar accuracy for this Dengue Virus dataset. This system produces the same results as that of Apriori algorithm. we can identify very clearly that Leucine (L), Phenylalanine (F), Lysine (K), Serine (S) is strongly associated with Glycine (G). In future this work can be compared with the FP-tree algorithm [21-23].

REFERENCES

1. Agrawal, R., T. Imielinksi and A. Swami, 1993. Database mining: a performance perspective, IEEE Transactions on Knowledge and Data Engineering, 5(6): 914-925.
2. Chen, M.S., J. Han and P.S. Yu, 1996. Data Mining: An Overview from a Database Perspective, IEEE Trans. Knowledge and Data Eng., pp: 866-883.
3. Agrawal, R., T. Imielinski and A. Swami, 1993. Mining Association rules between sets of items in large databases, In the Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93),Washington, USA, pp: 207-216.
4. Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules, in Proc. 20th Int. Conf. Very Large Data Bases, VLDB, edited by J.B. Bocca, M. Jarke and C. Zaniolo, Morgan Kaufmann, 12: 487-499.
5. Pei, M., E.D. Goodman and F. Punch, 2000. Feature Extraction using genetic algorithm, Case Center for Computer-Aided Engineering and Manufacturing W. Department of Computer Science.
6. Stuart, J. Russell and Peter Norvig, 2008. Artificial Intelligence: A Modern Approach.
7. Goldberg David, E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison Wesley, pp: 41.
8. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques, Morgan and Kaufmann, 2000.
9. Pujari, A.K., 2010. Data Mining Techniques, Universities Press. International Journal of Artificial Intelligence and Applications (IJAIA), 1(4): 143.
10. Anandhavalli, M., Suraj Kumar Sudhanshu, Ayush Kumar and M.K. Ghose, 2009. Optimized association rule mining using genetic algorithm, Advances in Information Mining, ISSN: 0975-3265, 1(2): 01-04.
11. Markus Hegland, 2005. The Apriori Algorithm-a Tutorial. CMA, Australian National University, WSPC/Lecture Notes Series, pp: 22-27.
12. Kazuo Sugihara, Measures for Performance Evaluation of Genetic Algorithms. Dept. of ICS, Univ. of Hwaii at Manoa.
13. LIU, B., W. HSU, S. CHEN and Y. MA, 2000. Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems.
14. David Beasley, *et al.,* 1993. An overview of genetic algorithms, Part 1 and 2, University Computing, 15(2/4): 58-69, 170-181.
15. GHOSH, A. and B. NATH, 2004. Multi-objective rule mining using genetic algorithms. Information Sciences, 163: 123-133.
16. Kerana Hanirex, D. and K.P. Kaliyamurthie, 2013. Finding The Dominating Amino Acids In Dengue Virus (Type-1): Study On Mining Frequent temsets, Int. J. Pharm. Bio. Sci., 4(3): (B)1246-1251.
17. Kumaravel, B. Anatha Barathi, 2013. Personalized image search using query expansion, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 15(12): 1736-1739.
18. Kumaravel, A. and R. Udayakumar, 2013. Web Portal Visits Patterns Predicted by Intuitionistic Fuzzy Approach, Indian Journal of Science and Technology, ISSN: 0974-6846, 6(5S): 4549-4553.
19. Kumaravel, A. and K. Rangarajan, 2013. Algorithm for Automation Specification for Exploring Dynamic Labyrinths, Indian Journal of Science and Technology, ISSN: 0974-6846, 6(5S): 4554-4559.
20. Kumaravel, A. and Oinam Nickson Meetei, 2013. An Application of Non-uniform Cellular Automata for Efficient Crytography, Indian Journal of Science and Technology, ISSN: 0974-6846, 6(5S): 4560-4566.
21. Pattanayak, Monalisa and P.L. Nayak, 2013. Green Synthesis of Gold Nanoparticles Using Elettaria cardamomum (ELAICHI) Aqueous Extract World Journal of Nano Science and Technology, 2(1): 01-05.

22. Chahataray, Rajashree and P.L. Nayak, 2013. Synthesis and Characterization of Conducting Polymers Multi Walled Carbon Nanotube-Chitosan Composites Coupled with Poly (P-Aminophenol) World Journal of Nano Science and Technology, 2(1): 18-25.

23. Parida, Umesh Kumar, S.K. Biswal, P.L. Nayak and B.K. Bindhani, 2013. Gold Nano Particles for Biomedical Applications World Journal of Nano Science and Technology, 2(1): 47-57.