CHAPTER 6
OPTIMIZED ASSOCIATION RULE MINING USING GENETIC ALGORITHM

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by major technologies such as massive data collection, powerful multiprocessor computers, data mining algorithms and that are now sufficiently mature [77].

6.1 Introduction

Association Rule Mining is one of the most commonly used algorithm in the research of data mining. These algorithms can be used for discovering hidden relationship between items. Given a user-specified threshold, also known as minimum support, the mining of association rules can discover the complete set of frequent patterns.

Association Rule Mining is also used for market basket analysis, which is the analysis of the itemset or products that can be analyzed after the customer purchasing the products from the shop. It is just like the analysis of the customers purchasing behavior. In OARM, the genetic algorithm is applied over the rules fetched from Apriori association rule mining. By using Genetic Algorithm, the system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. Since genetic algorithm is based on the greedy approach, the major advantage is that, in the discovery of prediction rules, they perform global search and its complexity is less compared to other algorithms.
6.2 Related Concepts

Association rule, apriori algorithm and genetic algorithm are discussed briefly in the further sub sections.

6.2.1 Association Rule

Association analysis is applicable in all the major application domains such as bioinformatics, geo informatics, data mining, web mining, medical diagnosis and scientific data analysis. The fundamental concepts of association rule mining are already discussed in brief in chapter 5. All the traditional association rule mining algorithms were developed to find positive associations between items. Positive associations refer to associations between items existing in transactions. In addition to the positive associations, negative associations can provide valuable information [78][82][84]. The formulae for Support and Confidence are also discussed in chapter 5.

The goal of mining association rules is to generate all possible rules that exceed specified minimum user-specified support and confidence thresholds.

6.2.2 Genetic Algorithm

A Genetic Algorithm (GA) is a procedure used to find approximate solutions to search problems through application of the principles of evolutionary biology. Genetic algorithms are typically implemented using computer simulations in which an optimization problem is specified. For this problem, members of a space of candidate solutions, called individuals, are represented using abstract representations called chromosomes [80, 81].GA is an iterative procedure that is appropriate for situations such as large and complex search space and optimization problems. General functions of Genetic Algorithms are Crossover, Selection, Replication, and Mutation. Genetic Algorithm is discussed in the next chapter 7, where again these concepts are used.

Fitness function is used during each iteration of the genetic algorithm to evaluate the quality of all the proposed solutions to the problems in the current population. The fitness function evaluates how good a single solution in a
population is, e.g. if one is trying to find for what \( x \)-value a function has its \( y \)-minimum with a Genetic algorithm, the fitness function for a unit might simply be the negative \( y \)-value (the smaller the value higher the fitness function).

The problem was formulated as follows:

Given a large database of item transactions, find the frequent itemsets, i.e. itemsets that occurs in at least a user-specified percentage of the database and also find the different patterns generated.

The performance of methods to mine frequent or maximal patterns depends on the length distribution of mined patterns. Here, algorithms will make groups of similar itemset collections to the database.

### 6.3 Experimental methodology

The project was run on Windows XP as operating system and the software used were JAVA (JDK 1.6.0), TCP/IP protocol and IDE Eclipse.

**Methodology**

The proposed method for generating association rule by genetic algorithm is as follows:

Step 1: Start

Step 2: Load a sample of records from the database that fits in the memory.

Step 3: Apply Apriori algorithm to find the frequent item sets with the minimum support. Let \( A \) is set of the frequent item set generated by Apriori algorithm.

Step 4: Set \( B = 0 \) where \( B \) is the output set, which contains the association rule.

Step 5: Input the termination condition of genetic algorithm.

Step 6: Represent each frequent item set of \( A \) as binary string using the combination of representation.

Step 7: Select the two members from the frequent item set using Roulette Wheel sampling method.

Step 8: Apply the crossover and mutation on the selected members to generate
the association rules.

Step 9: Find the fitness function for each rule $X \rightarrow Y$ and check the following condition.

Step 10: If (fitness function > min confidence)

Step 11: Set $B = B \cup \{X \rightarrow Y\}$

Step 12: If the desired number of generations is not completed, then go to Step 3.

Step 13: Stop

The system, developed using the above concepts, provides faster access to databases (Log files) through mining the data from a given huge collection of data, using which important knowledge were extracted. The minimum support considered for the analysis here was 20%.

Frequent itemsets generated by applying association rule alone take longer computation time for accessing information from huge log file (which keeps the information about past sessions). So by using genetic algorithm, the same results was generated in a more efficient manner. Thus the objective of the system to implement association rule mining of data using genetic algorithm to find all the frequent itemsets from the given dataset in an efficient manner and improve the performance of accessing information from databases (Log file) maintained at server machine leading to self optimization was achieved.

The data set used was that of a bakery sales which consists of entries in the form of a sparse vector representation:

Receipt# followed by item #'s that are on that receipt

The dataset is that of a bakery chain that has a menu of about 40 pastry items and 10 coffee drinks. It has a number of locations in West Coast states (California, Oregon, Arizona, Nevada). The dataset file describes the contents of the EXTENDED BAKERY dataset developed for CPE 466, Knowledge Discovery in Data course at Cal Poly. It contains information about one year worth of sales information for a couple of small bakery shops. The dataset contains information about the different store locations, the assortments of baked goods offered for sale
and the purchases made. This is the standard dataset available online for transaction analysis.

6.4 Results and interpretation

The following were the observations on applying the Optimized Apriori Rule Mining Algorithm with GA on a dataset of size 1000 and 5000. Tabulation of the values of different factors for these two sizes of dataset with respect to number of groups or clusters generated, range of number of items purchased, least number of receipts in a group, highest number of receipts in a group, and number of different patterns is shown below.

Table 6.1: Tabulation of the values of different factors for 2 sizes of dataset

<table>
<thead>
<tr>
<th>Factors</th>
<th>1000 receipts</th>
<th>5000 receipts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of gs generated</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>Range of number of items purchased</td>
<td>2 – 8</td>
<td>2- 35</td>
</tr>
<tr>
<td>Least number of receipts in a group</td>
<td>51</td>
<td>259</td>
</tr>
<tr>
<td>Highest number of receipts in a group</td>
<td>103</td>
<td>544</td>
</tr>
<tr>
<td>Number of different patterns</td>
<td>2097</td>
<td>10310</td>
</tr>
</tbody>
</table>

In table 6.2, the first column represents the group number. The second column represents the number of transactions in that respective group. The last column shows the items which are likely to be bought in a transaction belonging to that group. For example, in group 28, there are 51 transactions. In most of the transactions, items 7, 11 and 45 are present. It means if a customer buys item 7, it’s most likely that he/she will buy items 11 and 45 also. The system is also able to display the count of the number of groups formed with different size of datasets automatically.
Table 6.2: Case 1: Result of 1000 receipts dataset analysis

<table>
<thead>
<tr>
<th>Group No.</th>
<th>No. of Receipts</th>
<th>Most common item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>67</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>77</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>91</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>63</td>
<td>44</td>
</tr>
<tr>
<td>10</td>
<td>103</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>77</td>
<td>4,9</td>
</tr>
<tr>
<td>12</td>
<td>91</td>
<td>5,14,44</td>
</tr>
<tr>
<td>13</td>
<td>79</td>
<td>12,31,48</td>
</tr>
<tr>
<td>14</td>
<td>72</td>
<td>12,31,36,48</td>
</tr>
<tr>
<td>15</td>
<td>57</td>
<td>12,31,36,48</td>
</tr>
<tr>
<td>16</td>
<td>76</td>
<td>31,36,48</td>
</tr>
<tr>
<td>17</td>
<td>81</td>
<td>27,28</td>
</tr>
<tr>
<td>18</td>
<td>68</td>
<td>7,11</td>
</tr>
<tr>
<td>19</td>
<td>73</td>
<td>23,24,30,31</td>
</tr>
<tr>
<td>20</td>
<td>58</td>
<td>23,24,42</td>
</tr>
<tr>
<td>21</td>
<td>51</td>
<td>9,13</td>
</tr>
<tr>
<td>22</td>
<td>53</td>
<td>1,19</td>
</tr>
<tr>
<td>23</td>
<td>77</td>
<td>22</td>
</tr>
<tr>
<td>24</td>
<td>74</td>
<td>16</td>
</tr>
<tr>
<td>25</td>
<td>56</td>
<td>7,11,37,45</td>
</tr>
<tr>
<td>26</td>
<td>82</td>
<td>45</td>
</tr>
<tr>
<td>27</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td>28</td>
<td>51</td>
<td>7,11,45</td>
</tr>
<tr>
<td>29</td>
<td>69</td>
<td>47</td>
</tr>
<tr>
<td>30</td>
<td>62</td>
<td>28</td>
</tr>
<tr>
<td>31</td>
<td>83</td>
<td>0,2,46</td>
</tr>
<tr>
<td>32</td>
<td>59</td>
<td>42</td>
</tr>
<tr>
<td>33</td>
<td>72</td>
<td>33</td>
</tr>
<tr>
<td>34</td>
<td>77</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Study and analysis of autonomic components to improve the performance of some database applications

No. of receipts in individual groups for 1000 datasets

Figure 6.1: Graphical representation of number of groups Vs number of receipts in that group (for 1000 dataset)

No. of Receipts in individual groups for 5000 datasets

Figure 6.2: Graphical representation of number of groups Vs number of receipts in that group (for 5000 dataset)
Table 6.3: Case 2: Result of 5000 receipts dataset analysis

<table>
<thead>
<tr>
<th>Group Number</th>
<th>Number of data sets in each group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>416</td>
</tr>
<tr>
<td>2</td>
<td>477</td>
</tr>
<tr>
<td>3</td>
<td>432</td>
</tr>
<tr>
<td>4</td>
<td>368</td>
</tr>
<tr>
<td>5</td>
<td>342</td>
</tr>
<tr>
<td>6</td>
<td>259</td>
</tr>
<tr>
<td>7</td>
<td>337</td>
</tr>
<tr>
<td>8</td>
<td>415</td>
</tr>
<tr>
<td>9</td>
<td>414</td>
</tr>
<tr>
<td>10</td>
<td>443</td>
</tr>
<tr>
<td>11</td>
<td>297</td>
</tr>
<tr>
<td>12</td>
<td>288</td>
</tr>
<tr>
<td>13</td>
<td>328</td>
</tr>
<tr>
<td>14</td>
<td>375</td>
</tr>
<tr>
<td>15</td>
<td>544</td>
</tr>
<tr>
<td>16</td>
<td>392</td>
</tr>
<tr>
<td>17</td>
<td>372</td>
</tr>
<tr>
<td>18</td>
<td>296</td>
</tr>
<tr>
<td>19</td>
<td>324</td>
</tr>
<tr>
<td>20</td>
<td>313</td>
</tr>
<tr>
<td>21</td>
<td>311</td>
</tr>
<tr>
<td>22</td>
<td>270</td>
</tr>
<tr>
<td>23</td>
<td>326</td>
</tr>
<tr>
<td>24</td>
<td>380</td>
</tr>
<tr>
<td>25</td>
<td>364</td>
</tr>
<tr>
<td>26</td>
<td>296</td>
</tr>
<tr>
<td>27</td>
<td>346</td>
</tr>
<tr>
<td>28</td>
<td>290</td>
</tr>
<tr>
<td>29</td>
<td>324</td>
</tr>
<tr>
<td>30</td>
<td>299</td>
</tr>
<tr>
<td>31</td>
<td>443</td>
</tr>
<tr>
<td>32</td>
<td>324</td>
</tr>
<tr>
<td>33</td>
<td>358</td>
</tr>
<tr>
<td>34</td>
<td>324</td>
</tr>
<tr>
<td>35</td>
<td>326</td>
</tr>
</tbody>
</table>

The graph 6.1 and graph 6.2 indicates that the number of groups formed with dataset of size 1000 was 34, whereas the number of groups formed for the dataset
of size 5000 was 35. That is, a very nominal change in the number of groups formed was observed. However, the density of the groups increased drastically. It was found that the total number of items in group 1 was 416 out of 5000; total number of items in group 2 was 477 out of 5000 and so on.

This market basket analysis carried out together using association rule algorithm and genetic algorithm allows the analyst and decision makers to observe patterns associated with shopping and accordingly take required actions or steps either to retain or improvise it.

### 6.5 Snapshots

![Figure 6.3: Total number of groups formed for a dataset of size 1000](image)

Above figure 6.3 shows the number of groups formed for a dataset of size 1000 and the receipts grouped in this category. It was found that the item 1 was the most frequent item in this group.
Study and analysis of autonomic components to improve the performance of some database applications

Figure 6.5: Screenshot showing the number of transaction in group 24 for a dataset size of 5000

Figure 6.6: Screenshot showing the number of different patterns for a dataset size of 1000
6.6 Summary

Old system of Association rule mining has dealt with a challenging association rule mining problem of finding optimized Association rules. The frequent itemsets are generated using the Apriori association rule mining algorithm. The genetic algorithm has been applied on the generated frequent itemsets to generate the rules containing positive attributes, the negation of the attributes with the consequent part consisting of single attribute or more than one attribute.

The OARM approach also provides practical advantage over many existing techniques whose application requires customized and complex runtime environment.

This study shows that using apriori algorithm and genetic algorithm together, the system could be optimized with the available historical data. This information collected will be useful to identify the number of groups formed based on the items or products bought, buying pattern of the customers, frequency of buying a particular product, and also the most different pattern of purchase. Basically in this experiment carried out, optimization of the routine association rule mining algorithm is done with the help of genetic algorithm. This ability of the system to identify the groups of frequent itemsets, number of groups formed and also the group of different patterns based on the support value is very critical in decision making of the placing of the products on the shelf, identifying the demand of a particular product individually and in association with other products. This knowledge discovery from the available data in the receipts is very much useful in refining the policy decisions in an organization as well. This information also provides an insight of the fast moving item or product in the market.

Decision making is one of the major features in an autonomous system.