CHAPTER - 9

CONCLUSION AND FUTURE WORK

9.1. CONCLUSION

Among the three algorithms of Apriori, Predictive Apriori and Tertius, the best rules are always found by Predictive Apriori Algorithm (see table 3.5). It is independent of the number of instances and attributes. Suppose ‘D’ denotes the data set and ‘t’ denotes the time and ‘i’ denotes the items that are generated then, during the execution it has been observed that $D \alpha t$ i.e. the execution time of Apriori algorithm increases as the data set increases. It has also been observed that $D \alpha 1/i$ i.e. the number and size of frequent itemsets generated by strong association rule in Apriori decreases with the increase in datasets.

Out of the three algorithms used for classification (see table 4.4) NB Tree has the highest accuracy, followed by NB and MLP. The numerical experiments have revealed that MLP has the highest miscalculation rate and lowest precision value. The training time for the classifiers is the least with NB whereas the time taken to build the model is the least with NBTree. The analysis shows that NBTree and NB Classifiers are good solutions for classification problems with many instances since MLP results in maximum incorrect classifications. NBTree and NB Classifiers are useful for a faster and correct method of early diagnosis of breast cancer.

From the analysis of prediction (see table 5.4, 5.7, 5.9, 5.11) it is found that when the data set consists of non-categorical values JRip, PART and Decision Tree categorize the data nearly to the true class. All the methods yield the most consistent accurate result for all the algorithms. All the algorithms except ZeroR have a consistent sensitivity rate for all the methods. Decision Table has a consistent specificity rate of 95% for all methods except cross validation. ZeroR algorithm results in 0% specificity and 0 NPV which means that this algorithm has the highest probability of mis-classifications.
It is observed that the efficiency of a mining algorithm is found to be the function of many variables such as dataset consisting of huge historical information, the perspective of collected information, context in which predicted information will be used. The steps of algorithm, the models embedded in the algorithm and the statistical methods such as cross validation, percentage split used in prediction process also play an important role. Prediction model is not a linear model to the present case study. It is also inferred that reliable results can be obtained if the mentioned points are considered in the algorithm design.

Each of the categories of clustering algorithms has their own strengths and limitations and therefore suited for different tasks. The case study of clustering revealed that hierarchical clustering algorithms are the most efficient for the analysis of biological databases such as those involving the analysis of gene sequences. This is because it provides detailed exploratory analysis of the sequences and has minimal error even when the number of samples being analyzed is increased.

Applying Fuzzy logic on both ARM and Clustering techniques on the breast cancer data set shows that attribute age [40-99] plays a major role in generating the rules and forming clusters. The table shows that for a few attributes like ‘node-caps’, ‘irradiate’ ‘degree-malignant’, the percentage of their occurrence in each cluster varies. The attributes that take part in the determination of cancer are age [40-49][50-59], menopause with value being ‘premeno’, node-caps being ‘no’, degree of malignancy being ‘2’, breast quadrant being ‘left_up’ and irradiate being ‘no’. FARM algorithm works only for quantifiable data and non-categorical data since it implements K-means algorithm internally.

Although the primary causes of cancer are not yet known, there are a number of risk factors that have been identified and can therefore be fixed to particular classes. Generally, tumors can be malignant (cancerous) or benign (non-cancerous). In most cases, malignant tumors have rapid growth that often results in the destruction of normal tissues and their eventual spread to all parts of the body. The experiment results in the findings that age, menopause and degree of malignancy are some of the reasons for breast cancer.
The work done by extending Horspool algorithm reduces the need to store the large input of sequences during the entire execution of the program to match the subsequence.

The analysis of the prediction based on mode as an identifier predicts that attributes irradiat, inv-nodes, node-caps and menopause make the monotonically decreasing impact on the result of prediction which leads to deducing a classification rule.

The analysis of the second experiment based on extracting indicator attributes results in Mitoses and Bare_Nuclei being identified as the indicator attributes. The result obtained matched with the literature available on this subject that is mitotic index could stratify women into groups with high and low risk of recurrence. This result is also found to be true as per National Institutes of Health report gene's Position in the Nucleus can be used to distinguish Cancerous from Normal Breast Tissue.

The investigation helps in detecting diseases from large data sets through right selection of algorithms. This research addresses the necessity of improving the efficiency of data mining algorithms on large biological databases for early and precise detection of diseases. Comparative studies carried out on various data sets and algorithms show the efficiency of algorithms on each data set and data mining task. The algorithms derived in this process helps in achieving the objectives considered in this research work and are namely.

1. Extension of Horspool algorithm
2. Prediction based on quantifier attribute
3. Prediction based on identifier attribute

9.2. FUTURE WORK

The research carried out on large data sets related to various diseases reveals that all the algorithms do not give the same efficiency. It mainly depends on the data mining tasks. Further the efficiency and suitability of the algorithms depend on whether the data is categorical or non-categorical.
For association rule mining Predictive Apriori algorithm is found to be best. When the instances are large NB and NBTree classifiers are good solutions. However, NBTree is a better choice for classification. For predicting new information Decision Table algorithm is the best choice. While clustering hierarchical algorithms are the most efficient. FARM algorithm works only on non-categorical data. The extension of Horspool algorithm reduces the storage of large input of sequences and improves the total time taken to match the sub-sequence. Prediction done based on the modal values works for categorical data and indicator attributes works for non-categorical data.

The association mining algorithms can be modified to work on non-categorical data. The prediction model based on quantifier attribute and identifier attribute are very specific in nature. In the future they can be generalized so that it can work on any data set and give specific and best results. This investigation will help in detecting diseases from large data sets.