Association Technique in Data Mining and Its Applications

Harveen Buttar*, Rajneet Kaur**

*(Research Scholar, Department of Computer Science Engineering, SGGSWU, Fatehgarh Sahib, Punjab, India.)

**(Assistant Professor, Department of Computer Science Engineering, SGGSWU, Fatehgarh Sahib, Punjab, India.)

***Abstract*-- Data mining provides us with a variety of techniques for pattern analysis on large data such as association, clustering, segmentation and classification for better manipulation of data. This paper presents that how the data mining technique : association can be used in different areas. For instance, this technique helps the pharma firms to compete on lower costs while improving the quality of drug discovery and delivery methods. Also, this technique can be helpful in breast cancer diagnosis and prognosis. It shows that association rule mining algorithms can be used in the classification approach.**

***Keywords*-- Data Mining, Assocation, Drug Discovery, Breast Cancer.**

I. INTRODUCTION

Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. there is growing evidence that merging classification and association rule mining together can produce more efficient and accurate classification systems than traditional classification techniques [1].

Classification and association rule discovery are similar except that classification involves prediction of one attribute, i.e., the class, while association rule discovery can predict any attribute in the data set. In the last few years, a new approach that integrates association rule mining with classification has emerged [1, 2, 3].

II. ASSOCIATION RULE MINING

Since the presentation of association rule mining by aggarwal, imielinski and swami in their paper " Mining association rules between sets of items in large databases " in 1993 [4], this area remained the most active research areas in machine learning and knowledge discovery. Presently, association rule mining is one of the most important task in data mining. It is considered a strong tool for market basket analysis that aims to investigate the shopping behavior of customers in hoping to find regularities [4]. In finding association rules, one tries to find group of items that are frequently sold together in order to infer items from the presence of other items in the customer's shopping cart. For instance, an association rule may state that " 80% of customers who buy diaper and ice also buy cereal ". This kind of information may be beneficial and can be used for strategic decisions like items shelving, target marketing, sale promotions and discount strategies.

Association rules is a valuable tool that has been used widely in various industries like supermarkets, mail ordering, telemarketing, insurance fraud, and many other applications where finding regularities are targeted. The task of association rules mining over a market basket has been described in [4], formally, let D be a database of sales transactions, and let I={i1,i2,.....,im} be a set of binary literals called items. A transaction T in D contains a set of items called itemset, such that $\boldsymbol{T \subseteq I}$. Generally, the number of items in an itemset is called length of an itemset. Itemsets that have a length k are denoted by k-itemsets. Each itemset is associated with a statistical threshold named support. The support of the itemset is the number of transactions in D that contains the itemset. an association rule is an expression $X \Rightarrow Y$,where $X,Y \subseteq I$ are two sets of items and $X \cap Y = \varphi$. X is called the antecedent, and Y is called the consequent of the association rule. An association rule $X \Rightarrow Y$ has a measure of goodness named confidence, which can be defined as,

the probability a transaction contains Y given that it contains X, and is given as support(XY)/support(X).

Given the transactional database D, the association rule problem is to find all the rules that have a support and confidence greater than certain user specified thresholds, denoted by minsupp and minconf, respectively.

The problem of generating all association rules from a transactional database can be decomposed into two sub-problems :

1. The generation of all itemsets with support greater than the minsupp. These itemsets are called frequent itemsets. All other itemsets are called infrequent.

2. For each frequent itemset generated in step 1, generate all rules that pass minconf threshold. For example, if item XYZ is frequent, then we might evaluate the confidence of the rules $XY \Rightarrow Z$, $XZ \Rightarrow Y$ and $YZ \Rightarrow X$.

Table 1 : Transactional Database

Transaction Id	Item	Time
I1	bread, milk, juice	10:12
I2	bread, juice, milk	12:13
I3	milk, ice, bread, juice	13:22
I4	bread, eggs, milk	13:26
I5	ice, basket, bread, juice	15:11

For clarity, consider for example the database shown in table 1, and let minsupp and minconf be 0.70 and 1.0, respectively. The frequent itemsets in Table 1 are {bread}, {milk}, {juice}, {bread, milk} and {bread,juice}. The association rule that pass minconf among these frequent itemsets are *milk* ⇒ *bread* and *juice* ⇒ *bread.*

While the second step of association rule discovery that involves generation of the rules is considerably a straightforward problem given that the frequent itemsets and their *support* are known [4, 8, 9, 10]. The first step of finding frequent itemsets is relatively a resource consuming problem that requires extensive computation and large resource capacity especially if the size of the database and the itemsets are large [4, 11, 12]. Generally, for a number of distinct items *m* in a customer transaction database *D,* there are 2^m possible number of itemsets. Consider for example a grocery store that contains 2100 different distinct items. Then there are 2^{2100} possible different combinations of potential frequent itemsets, known by candidate itemsets, in which some of them do not appear even once in the database, and thus usually only a small subset of this large number of candidate itemsets is frequent. This problem has extensively being investigated in the last decade for the purpose of improving the performance of candidate itemsets generation [12, 11, 13, 14].

One of the first algorithms that has significant improvements over the previous association rule algorithms is the Apriori algorithm[15]. The Apriori algorithm presents a new key property named the "downward-closure" of the support, which states that if an itemset passes the *minsupp* then all of its subset must also pass the *minsupp.* This means that any subset of a frequent itemset have to be frequent, where else, any superset of infrequent itemset must be infrequent. Most of the classic association rule algorithms which have been developed after the Apriori algorithm such as [11, 12] have used this property in the first step of association rules discovery. Those algorithms are referred to as the Apriori-like algorithms or techniques.

III. ASSOCIATION TECHNIQUE FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE

Breast cancer has become the primary reason of death in women in developed countries. The most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The various methodology used are given below.

A. Decision Trees

Decision Trees treatments for breast cancer are, local and systematic. Surgery and radiation are local treatments. Decision Tree Learning is one of the most widely used and practical methods for classification. In this method, learned trees can be represented as a set of if-then rules that improve human readability. Decision trees are very simple to understand and interpret by domain experts. A decision tree consists of nodes that have exactly one incoming edge, except the root node that has no incoming edges. A node with outgoing edges is an internal node, while the other nodes are called leaves or terminal nodes or decision nodes. The experiments, is conducted using Weka J48, C4.5 decision tree is generated . Several parameters were tested such as the confidence factor 6 O. used for pruning, whether to use binary splits or not, whether to prune the tree or not and the minimum number of instances per leaf. For each different combination of parameters, the average classification accuracy of the 10 folds is saved. The best combination of parameters is selected, with higher average classification accuracy on

10-fold cross validation.[5] The final model is shown in Figure 1.

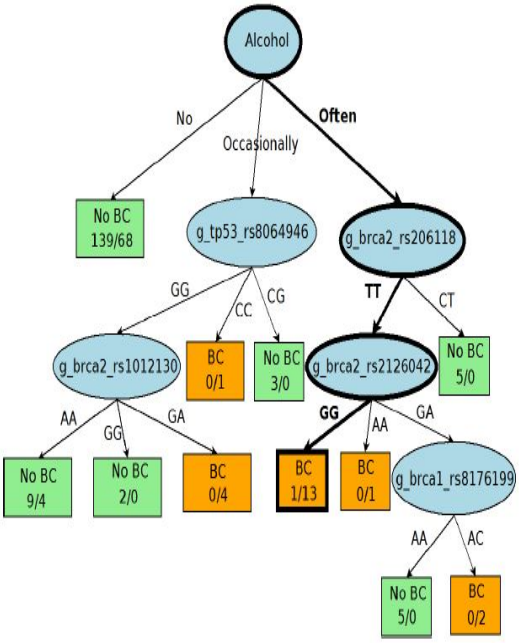

Fig 1 : Decision Tree Model

B. Digital Mammography Classification Using Association Rule Mining and ANN

In [6] the authors have performed some experiments for tumor detection in digital mammography. In this, different data mining techniques, neural networks and association rule mining, have been used for anomaly detection and classification. The experiments conducted, demonstrate the use and effectiveness of association rule mining in image categorization.

The real medical images used in the experiments were taken from the Mammographic Image Analysis Society (MIAS). It consists of 322 images, corresponding to three categories: normal, benign and malign. There were 208 normal images, 63 benign and 51 malign, which are considered abnormal. The abnormal cases are further divided in six categories: micro calcification, circumscribed masses, speculated masses, ill-defined masses, architectural distortion and asymmetry.

Figure 2 shows an overview of the categorization process adopted for both systems. The first step is represented by the image acquisition and image enhancement, followed by feature extraction. The last one is the classification.

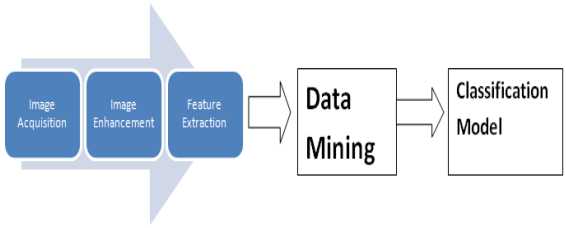

Fig 2 : Image Categorization Process

Mammograms [5] are the images but difficult to interpret, and a preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable. Two Image Enhancement techniques: a cropping operation and an image enhancement has been performed before feature extraction. After cropping and enhancing of the images, which basically represents the data cleaning phase, features relevant to the classification are extracted from the cleaned images.

The existing features are:
a) The type of the tissue (dense, fatty and fatty glandular) ;
b) The position of the breast: either left or right.

The above extracted features are computed over smaller windows of the original image. The original image is split in four parts as is shown in Figure 3 . These four regions are again splitted in other four parts to obtain more accurate extraction of the features and for the further investigation of the localization. For the sixteen sub-parts of the original image the statistical parameters were computed.

The four regions of the first division, and then, for each of the areas is further divided in four. To automatically categorize medical images on real mammograms two data mining techniques, association rule mining and neural networks is used.

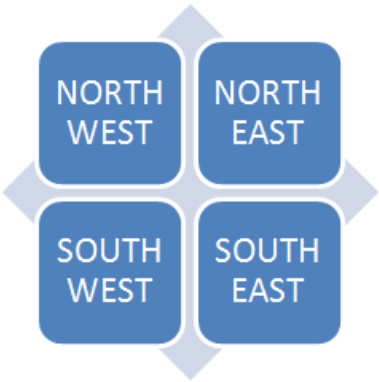

Fig 3

IV. ASSOCIATION RULE BASED CLASSIFIER

Association rule mining aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T1,\ldots., Tn\}$ and a set of items $I = \{i1, \ldots., im\}$ such that any transaction T in D is a set of items in I, an association rule is an implication of the form A goes to B where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence. Association rules mining is a two-step process, in the first step frequent item-sets are generated (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived from the frequent item-sets obtained in the first step.

The apriori algorithm [5] is used in order to discover association rules among the features extracted. After all the features are merged and put in the transactional database, the next step is applying the apriori algorithm for finding the association rules in the database constrained as described above with the antecedent being the features and the consequent being the category. Once the association rules are found, they are used to construct a classification system that categorizes the mammograms as normal, malign or benign. The most delicate part of the classification with association rule mining is the construction of the classifier itself. In the training phase, the apriori algorithm was applied on the training data to extract association rules. The support was set to 10% and the confidence to 0%. The success rate for association rule classifier was 69.11% on average.

V. AAPLICATION IN PHARMACEUTICAL INDUSTRY

Most healthcare institutions lack the appropriate information systems to produce reliable reports with respect to other information than purely financial and volume related statements (Prins & Stegwee, 2000). The management of pharma industry starts to recognize the relevance of the definition of drugs and products in relation to management information. In the turmoil between costs, care-results and patient satisfaction the right balance is needed and can be found in upcoming information and Communication technology.[7]

Given the size of the databases being queried, there is likely to be a trade-off in accuracy of information and processing time. Sampling techniques and tests of significance may be satisfactory to identify some of the more common relationships; however, uncommon relationships may require substantial search time.[7] The thoroughness of the search depends on the importance of the query (e.g., life threatening vs. "curious to know"), the indexing structures used, and the level of detail supplied in the query. Of course, the real data mining challenge comes when the user supplies only a minimal amount of information.

A user-interface may be designed to accept all kinds of information from the user (e.g., weight, sex, age, foods consumed, reactions reported, dosage, length of usage). Then, based upon the information in the databases and the relevant data entered by the user, a list of warnings or known reactions (accompanied by probabilities) should be reported. Note that user profiles can contain large amounts of information, and efficient and effective data mining tools need to be developed to probe the databases for relevant information. Secondly, the patient's (anonymous) profile should be recorded along with any adverse reactions reported by the patient, so that future correlations can be reported. Over time, the databases will become much larger, and interaction data for existing medicines will become more complete.

There are in general three stages of drug development namely finding of new drugs, development tests and predicts drug behavior, clinical trials test the drug in humans and commercialization takes drug and sells it to likely consumers (doctors and patients).

VI. CONCLUSION

Association is one of the best technique known in data mining. This paper demonstrates the use of association rule mining in classification approach which is found very useful in the pharmaceutical industry and in the diagnosis and prognosis of cancer. In future, a more improved classifier can be developed based on the association technique which may provide much better results.

VII. REFERENCES

[1] Liu, B., Hsu, W., and Ma, Y. (1998). Integrating Classification and association rule mining. In KDD '98, New York, NY, Aug. 1998.

[2] *Thabtah, F., Cowling, P. and Peng, Y. H. (2004). MMAC: A New Multi-Class, Multi-Label Associative Classification Approach. Fourth IEEE International Conference on Data Mining (ICDM'04).*

[3] *Li, W., Han, J., and Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple-class association rule. In ICDM'01, pp. 369-376, San Jose, CA.*

[4] Agrawal, R., Amielinski, T., and Swami, A. (1993). *Mining association rule between sets of items in large databases*. In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC, May 26-28.

[5] Shweta Kharya," Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease" , International Journal of Computer

Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[6] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coma, .Application of Data Mining Techniques for Medical Image Classification.Proceeding of second International worshop on Mutimedia data mining(MDM/KDD'2001),in conjuction with ACM SIGKDD conference.SAN FRANCISCO,USA,AUG 26,2001

[7] Journal of Theoretical and Applied Information Technology (2005), Applications of Data Mining Techniques in Pharmaceutical Industry.

[8] Agrawal, R. and Srikant, R. (1994). *Fast algorithms for mining association rule*. *Proceedings of the 20th International Conference on Very Large Data Bases***.** pp. 487 - 499.

[9] Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). *Algorithms for association rule mining — a general survey and comparison*. SIGKDD Explor. Newsl. 2, 1 (Jun. 2000), 58-64. DOI= http://doi.acm.org/10.1145/360402.360421

[10] Li, J., Zhang, X., Dong, G., Ramamohanaro, K. and Sun, Q. (1999). *Efficient mining of high confidence rules without support thresholds.* A book chapter in the book "Principles of Data Mining and Knowledge Discovery". LNCS.

[11] Park, J. S., Chen, M., and Yu, P. S. (1995). *An effective hash-based algorithm for mining association rules*. SIGMOD Rec. 24, 2 (May.1995), 175-186. DOI= http://doi.acm.org/10.1145/568271.223813

[12] Blackmore, K. and Bossomaier, T. J. (2003). *Comparison of See5 and J48.PART Algorithms for Missing Persons Profiling*. Technical report. Charles Sturt University, Australia.

[13] Han, J., Pei, J., and Yin, Y. (2000) Mining frequent patterns without candidate generation.. *2000 ACM SIGMOD Intl. Conference on Management of Data*.

[14] Dong, G., Li, J. (1999). *Efficient mining of frequent patterns: Discovering trends and differences*. In Proceeding of SIGKDD 1999, San Diego, California.

[15] Agrawal, R. and Srikant, R. (1994). *Fast algorithms for mining association rule*. *Proceedings of the 20th International Conference on Very Large Data Bases***.** pp. 487 - 499.