International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Aug-2016 www.irjet.net p-ISSN: 2395-0072

Applying Association Rules and Decision Tree Algorithms with Tumor Diagnosis Data

Alaa Khalaf Hamoud

Lecturer Assist. , Dept. of Information Technology, College of Computer Science and Information Technology, Basra University, Iraq.

Abstract - *In the last decade, Tumor diagnosis data received a lot of attention and researches due to its importance. The development of the hardware and software technologies led to necessity for tools not to store them but also to analyze these medical records and find results to support decisions. Association rules and decision tree are interested mining algorithms which can be used to find and explore relations between attributes in a data set. In this paper, association rules and decision tree algorithms are applied with tumor data set in order to get analytical results to support medical decisions. The results can be depended in early detection of tumor rather to provide second opinion for clinicians. Data mining algorithms are applied on 7544 medical diagnosis cases from more than 10 medical centers in Basra, Iraq.*

***Key Words*: Data Mining, Medical Data Set, Decision Tree Algorithm, Association rules Algorithm**.

1.INTRODUCTION

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce human readable classification rules and easy to interpret than other classification methods [6].

Data mining tools used for classification, machine learning, artificial intelligence rather than a prior diagnosis of disease infection. Data mining tools can be used as essential tools to get results which can be used to support decisions. The clarity and accuracy of results are main reasons for choosing decision tree algorithm in this work. First of all, tumor diagnosis paper records are collected from more than ten medical centres in Basra, Iraq. The first stage, these medical records transformed into electronic medical records (EMR). This stage makes the data entry took more than two months in the manual input of medical data. SQL Server data tool 2012 used to apply decision tree and association rules algorithms on tumor diagnosis records.

The paper organized as followed: the second section of the paper views the related works. Section 3 explains the concept of data mining and it definition while section 4 views how decision tree algorithm works. Section 5 and 6 view the explanation of association rule algorithm and Apriori algorithm which depended on by association rules algorithm. Section 7 list the experimental results of both decision tree and association rules .Conclusion and future work listed in the last section.

2. Related Works

In [3] a hybrid approach, Classification and regression trees (CART) classifier with feature selection and bagging technique has been considered to evaluate the performance in terms of accuracy and time for classification of various breast cancer datasets.

In [5] use Data Mining techniques for the data analysis, data accessing and knowledge discovery procedure to show experimentally and practically that how consistent, able and fast are these techniques for the study in the particular field? A solid mathematical threshold (0 to 1) is set to analyze the data. The obtained outcome will be tested by applying the approach to the databases, data warehouses and any data storage of different sizes with different entry values. The results shaped will be of different level from short to the largest sets of tuple. By this, we may take the results formed for different use e.g. Patient investigation, frequency of different disease.

In [7] a novel computer aided diagnosis (CADs) system is described using data mining with decision tree for classification to increase the level of diagnostics confidence and to provide a second opinion for physician.

3. Data Mining

Data mining can help reduce information overload and improve decision making. This is achieved by extracting and refining useful knowledge through a process of searching for relationships and patterns from the extensive data collected by organizations. The extracted information is used to predict, classify, model, and summarize the data being mined. Data mining technologies, such as rule induction, neural networks, genetic algorithms, fuzzy logic and rough sets, are used for classification and pattern recognition in many industries[10].

Data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. So Data mining is

the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. [4].

Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. The field is young, dynamic, and promising. Data mining has and will continue to make great strides in our journey from the data age toward the coming information age [2].

4. Decision Tree Algorithm

A "divide-and-conquer" approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree [1]. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [2].

A node cannot be further divided if the data records in the data set at this node have the same value of the target variable. Such a node becomes a leaf node in the decision tree. Except the root node and leaf nodes, all other nodes in the decision trees are called internal nodes. The decision tree can classify a data record by passing the data record through the decision tree using the attribute values in the data record. [4].

Decision tree classification technique is performed in two phases [8]: Tree building and Tree pruning. Tree building is done in top-down manner. During this phase that the tree is recursively partitioned till all the data items belong to the same class label. It is very tedious tasking and computationally intensive as the training data set is traversed repeatedly. Tree pruning is done in bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set). Over-fitting in decision tree algorithm is the cause of misclassification error.

5. Association Rules Algorithm

Let $I=I_1, I_2, \ldots, I_m$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain $X \cup Y$ to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule $X \Rightarrow Y$ is 80%, it means that 80% of the transactions that contain X also contain Y together [9]

Association rules are really no different from classification rules except that they can predict any attribute, not just the class, and this gives them the freedom to predict combinations of attributes too. Also, association rules are not intended to be used together as a set, as classification rules are. Different association rules express different regularities that underlie the dataset, and they generally predict different things. Because so many different association rules can be derived from even a tiny dataset, interest is restricted to those that apply to a reasonably large number of instances and have a reasonably high accuracy on the instances to which they apply to [1].

6. Apriori Algorithm

The Apriori algorithm (Agrawal and Srikant, 1994) provides an efficient procedure of generating frequent item sets by considering that an item set can be a frequent item set only if all of its subsets are frequent item sets. Table (1) gives the steps of the Apriori algorithm for a given data set D [9].

Table (1): Apriori Algorithm Steps

Step	Description of the Step
1	F_1 = {frequent one-item sets}
2	$i = 1$
3	while $F_i \neq \emptyset$
4	$i = i + 1$
5	$C_i = \{\{x_1, \ldots, x_{i-2}, x_{i-1}, x_i\} \mid \{x_1, \ldots, x_{i-2}, x_{i-1}\} \in F_{i-1}$ and $\{x_1, \ldots, x_{i-2}, x_i\} \in F_{i-1}\}$
6	for all data records $S \in D$
7	for all candidate sets $C \in C_i$
8	if $S \supseteq C$
9	C.count = C.count + 1
10	$F_i = \{C \mid C \in C_i$ and C.count $\geq$ minimum support$\}$
11	return all $F_j, j = 1, \ldots, i-1$

7. Experimental Results

As mentioned before, the tumor diagnosis data are collected from more than ten different medical centers and took more than two months to convert them from paper-based records into electronic medical records as excel and access database.

Both decision tree and association rule algorithms can be used for classification and regression. Decision tree algorithm used for classification and association rule algorithm used for prediction.

7.1 Decision Tree

Decision tree algorithm is one of the most important data mining tools which used for classification and regression. Basically, the algorithm builds the tree based on testing states by using if-then rule and produces nodes. Starting from the root node, the algorithm tests the states based on input column values and constructs the nodes. The nodes can be either internal nodes which can be spilled into other nodes or leaf nodes which can be used by decision makers due to their values. The resulting graph from this algorithm is easy to understand by the users.

The result decision tree is implemented by using Microsoft Decision Tree algorithm which uses discrete and continuous data types as input values. The Input columns of decision tree algorithm are diagnosis column and district column. The decision tree classifies the data based on the number of tumor diagnosis cases and constructs the leaf nodes based on predict column (gender).

The figure (1) shows the decision tree of tumor cases where each leaf node (F and M) represent the number of female and male infections for specific tumor case.

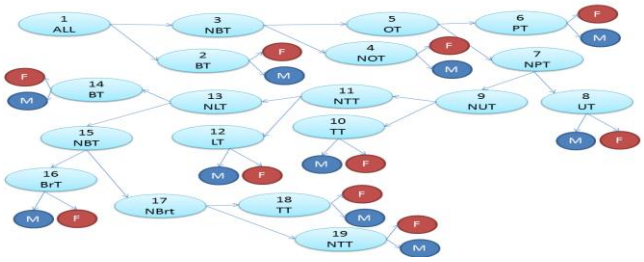

Figure (1): Decision Tree Model

Since the resulting graph from SQL Server Data tool is too large so it been converted into the graph in the figure (1). The root node started with all tumor cases (7544 cases). The rule used to divide cases based on a number of diagnoses for each case classified by gender. So, the first node is divided into BT(Breast Tumor) and NBT(non-Breast Tumor). BT is divided into F(Female with 1882 case registered) and M (Male with 66 cases registered) NBT (5596 cases) is divided into NOT(non-Ovarian Tumor) and OT(Ovarian Tumor) which classified into F(Female with 183 cases) and M (Male with6 case). Table (2) views the other details related to each node in the tree graph.

Table (2): Decision Tree nodes Values

Node number	Short-cut	Tumor	Female	Male	Unknown	Total
1	A	ALL	4258	3284	2	7544
2	BT	Breast-Tumor	1882	66	0	1948
3	NBT	Non-Breast-Tumor	2376	3218	2	5596
4	OT	Ovarian-Tumor	183	6	0	189
5	NOT	Non-Ovarian-Tumor	2193	3212	2	5407
6	PT	Prostat-Tumor	4	172	0	179
7	NPT	Non-Prostat-Tumor	2189	3040	2	5231
8	UT	Uterus-Tumor	96	4	0	100
9	NUT	Non-Uterus-Tumor	2093	3036	2	5131
10	TT	Testes-Tumor	3	53	0	57
11	NTT	Non-Testes-Tumor	2090	2982	2	5074
12	LT	Lung-Tumor	169	395	0	564
13	NLT	Non-Lung-Tumor	1921	2587	2	4510
14	BT	Bladder-Tumor	103	260	0	363
15	NBT	Non-Bladder-Tumor	1818	2327	2	4147
16	BrT	Bronchitis-Tumor	27	101	0	128
17	NBrt	Non-Bronchitis-Tumor	1791	2226	2	4019
18	TT	Thyroid-Tumor	21	48	0	69
19	NTT	Non-Thyroid-Tumor	1743	2205	2	3950

7.2 Association Rules

Association rule algorithm is the implementation of the well-known Apriori algorithm. Microsoft Association Rule Algorithm needs candidate's columns to apply Apriori algorithm in order to produce rules. The below table (3) shows the candidates columns.

Table (3): Candidates Columns of Association Rule Algorithm

Column	Type
Seq	Key
Diagnosis	Input
Gender	Input
Province	Input
District	Input
Age	Predict

The intended result of applying this algorithm is to produce rules that find the relationship between Input column and Predict column. In other words, the goal of making Age column as predictable is to find the target age or age range which has been infected by specific tumor and carry the attributes (gender, district, and province). The parameters of Microsoft Association Rules Algorithm are set as shown in the table (4).

Table (4): Association Rule Algorithm Parameters values

Parameter	Value
MAXIMUM_ITEMSET_COUNT	1026
MAXIMUM_ITEMSET_SIZE	3
MAXIMUM_SUPPORT	0.15
MINIMUM_ITEMSET_SIZE	2
MINIMUM_PROBABILITY	0.40
MINIMUM_SUPPORT	15
MINIMUM_ROWS	20000

The parameters are set to the values to get more reliable rules with high confidence and here the confidence is the probability. The maximum itemset count set to 1024 to hold all result item sets, the size of itemset represents the number of items (cases) in the item set. Increasing item set size may reduce the number of itemsets and then reduce the accurate rules. The probability is set high to filter out the unwanted and repeated rules.

Table (5): Association Rules

Pro...	Importance	Rule
0.521	0.373	Distinct = Maakal, Dignosis = Breast.Tumor -> Age = 25.7890020352 - 42.0320030208
0.517	0.268	Distinct = Qurna, Dignosis = Breast.Tumor -> Age = 42.0320030208 - 55.142064256
0.516	0.270	Distinct = UNKNOWN, Dignosis = Breast.Tumor -> Age = 42.0320030208 - 55.142064256
0.500	0.253	Distinct = Kiblah, Dignosis = Breast.Tumor -> Age = 42.0320030208 - 55.142064256
0.440	0.288	Dignosis = Breast.Tumor, Gender = F -> Age = 42.0320030208 - 55.142064256
0.434	0.283	Dignosis = Breast.Tumor -> Age = 42.0320030208 - 55.142064256
0.428	0.257	Dignosis = Breast.Tumor, Province = Basrah -> Age = 42.0320030208 - 55.142064256
0.400	0.262	Distinct = Midayna, Dignosis = Breast.Tumor -> Age = 25.7890020352 - 42.0320030208

The result rules shown in the table (5), the rules are filtered based on Breast Tumor Diagnosis to target this Tumor as candidate case. The predictable age range (25.7-40 years) in Maakal district has the higher probability of infection followed by age range (42-55 years) in Qurna.

3. CONCLUSIONS

Data mining tools often used in clinical path due to its ability to discover hidden pattern and to support decisions. Decision Tree and Association Rules algorithms mostly used to analyze association and provide the second opinion for the clinicians and decision makers. The result classification graph of Decision Tree can be considered as a roadmap for all registered tumor cases in Basra province in Iraq. It can be used to analyze the tumor behavior all around the province. While the association rules predict the ages for specific cases (Breast Tumor in the work) and can be filtered to another tumor to find easily the association rules.

This work can be depended as a platform to build online application decision support system which gives accurate analytical results. It also can be depended to study the behaviors of tumors and make precautionary measures to avoid the infections.

REFERENCES

[1] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[2] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[3] Lavanya, D., and K. Usha Rani. "Ensemble decision tree classifier for breast cancer data." International Journal of Information Technology Convergence and Services 2.1 (2012): 17.

[4] Ye, Nong. Data mining: theories, algorithms, and examples. CRC Press, 2013.

[5] Irshad ullah "Data Mining Algorithms And Medical Sciences." International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010.

[6] Lavanya, D., and K. Usha Rani. "Performance evaluation of decision tree classifiers on medical datasets." International Journal of Computer Applications (0975–8887 (, Volume 26–No. 4, 1-4 (2011).
[7] Kuo, Wen-Jia, et al. "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images." Breast cancer research and treatment66.1 (2001): 51-57.
[8] Anyanwu, Matthew N., and Sajjan G. Shiva. "Comparative analysis of serial decision tree classification algorithms." International Journal of Computer Science and Security 3.3 (2009): 230-240.
[9] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." GESTS International Transactions on Computer Science and Engineering 32.1 (2006): 71-82.
[10] Zhu, Dan. "Analytical Competition for Managing Customer Relations." (2009): 25-30.