K A B K Wijeratne – s15219

CS4104 – Data Analytics – Assignment 2 – Association Rule Mining

Objectives

Breast cancer remains one of the most common and deadly cancers affecting women globally. This study focuses on the domain of breast cancer diagnosis and classification, with an emphasis on early detection. Early detection is crucial for choosing appropriate treatment protocols and improving survival rates. While mammography is the primary technology for breast cancer screening, traditional diagnostic methods heavily depend on the expertise of doctors, making them susceptible to limitations and human error.

Recent advancements in science and technology, particularly in healthcare, have led to the collection of vast amounts of data on tumor characteristics, which hold the potential for uncovering crucial insights. Association rule mining, a data mining technique, has emerged as a valuable tool in this context. It allows the extraction of vital yet hidden patterns from large datasets, transforming raw data into meaningful and actionable knowledge.

By applying association rule mining to breast cancer data, we can discover high-level patterns that may not be immediately evident. These patterns can reveal important relationships between different tumor features and their association with cancer stages (benign or malignant). The use of these patterns can significantly improve the accuracy and speed of cancer-stage diagnosis, leading to more effective treatment decisions and ultimately enhancing patient survival rates. Consequently, this approach not only complements traditional diagnostic methods but also reduces reliance on human expertise, thereby minimizing errors and increasing the overall efficiency of breast cancer treatment.

Data set description

For this study, we utilized the Wisconsin Breast Cancer Dataset (WBCD), originally collected by Wolberg and Mangasarian at the University of Wisconsin-Madison Hospitals in 1990. This dataset, which can be accessed from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original), is a well-established resource in breast cancer research and analysis.

The WBCD comprises 699 instances, each representing a patient, with features derived from digitized images of fine needle aspirates (FNAs) of breast masses. These features include clump thickness, uniformity of cell size, uniformity of cell shape, and other critical attributes essential for distinguishing between benign and malignant tumors. This dataset is widely used in machine learning and data mining to develop predictive models for breast cancer diagnosis based on these tumor characteristics. An attribute description of the Wisconsin Breast Cancer Dataset, as it appears before preprocessing, is provided in Table 1 below.

Table 1: Attribute description of the Wisconsin Breast Cancer Dataset, as it appears before preprocessing

Variable name	Role	Type	Description	Missing
Sample_code_number	ID	Categorical		No
Clump_thickness	Feature	Numeric		No
Uniformity_of_cell_size	Feature	Numeric		No
Uniformity_of_cell_shape	Feature	Numeric		No
Marginal_adhesion	Feature	Numeric		No
Single_epithelial_cell_size	Feature	Numeric		No
Bare_nuclei	Feature	Numeric		yes
Bland_chromatin	Feature	Numeric		No
Normal_nucleoli	Feature	Numeric		No
Mitoses	Feature	Numeric		No
Class	Target	Binary	2=benign, 4=malignant	No

The data preprocessing was carried out as follows:

1. *Adding column names and Converting to an ARFF file*

The downloaded breast cancer dataset, initially in CSV format, was first saved with the appropriate column headers manually added using Excel, as specified in the repository. Subsequently, the dataset was loaded into Weka (Waikato Environment for Knowledge Analysis), an open-source machine learning software. It was then saved in ARFF format to meet the requirements for further analysis. After loading the ARFF file into Weka (Explorer > Choose File), the dataset contained 699 instances and 11 attributes (Figure 1).

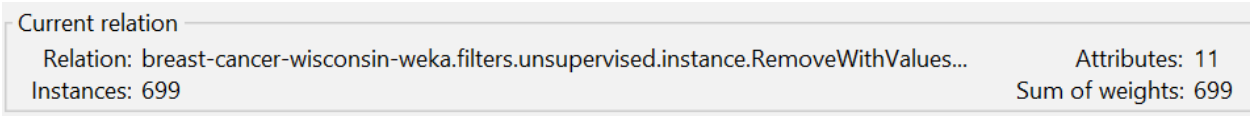

Figure 1: Dataset summary after loading to WEKA

2. *Removing Null values*

Next, null values in the dataset were removed using the RemoveWithValues filter in Weka. The parameters for the filter were set as follows: the attributeIndices field was configured to "7" to target the 'Bare Nuclei' column, where missing values were found. The matchMissingValues field was set to true to ensure that any instance with a missing value in the specified column would be considered a match and thus removed. After applying the filter, the dataset summary showed 689 instances, indicating that 10 instances with missing values had been removed (Figure 2).

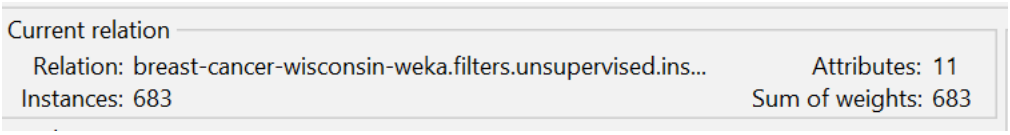

Figure 2: Dataset summary after removing missing values

3. *Removing Duplicates*

The data was saved after each preprocessing step to ensure backups. Duplicates, where all attribute values were identical, were removed using the weka.filters.unsupervised.instance.RemoveDuplicates filter. After this step, 675 instances remained (Figure 3). Some instances had the same sample code number but different attribute values; these records were not removed, assuming they represented errors in the sample code

number. However, removing records with identical attribute values is a crucial step in preprocessing to ensure the accuracy of the rule-mining process.

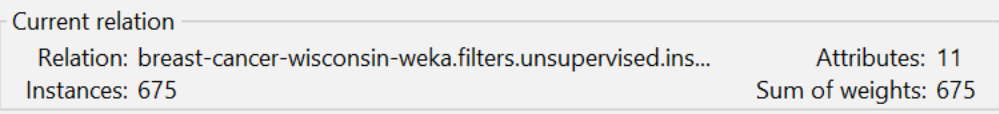

Figure 3: The dataset summary after removing the duplicates (for all attributes)

4. *Dropping the sample_code_number column*

Next, the first column, sample_code_number column was dropped using the weka.filters.unsupervised.attribute.Remove filter. The attributeIndices filed was set to 1 to specify the attribute that had to be removed. This reduces the amount of data needed to be processed and ensures that only meaningful attributes are considered in subsequent steps.

5. *Discretization of the dataset*

Next, the dataset was discretized to convert the numeric attributes to nominal attributes by dividing the range of the numeric values into a set number of intervals or bins. Each bin represents a range of values, and these bins are treated as categorical values. This was done using the weka.filters.unsupervised.attribute.Discretize filter and the number of bins were chosen by making the discretization supervised. This was done by setting the findNumBins attribute to True, which ensures that the program finds the optimal number of bins using the leave-one-out method. The attributeIndices field was set to 'first-last' to apply the filter to all columns. Discretization is essential for transforming the data into a categorical format that is suitable for association rule mining algorithms.

6. *Normalization – was not used here*

Normalization adjusts the scale of the data so that each feature contributes equally to the analysis. In association rule mining, this is important because it prevents features with larger ranges from dominating the rules. Normalization helps in maintaining the balance and comparability between different attributes, leading to more meaningful and accurate rules. If the dataset still contains numerical values after discretization, these values should be normalized so that all features contribute equally to the analysis. This step is less critical for purely categorical data hence it was not performed for the dataset as all attributes except the target class variable turned into nominal

type after discretization. Further, the attributes already had data within a scale of 1-10 hence it was unnecessary to normalize the data again as all the data belongs to a similar range.

7. Turning the target variable (class) to a Binary type.

The class label attribute values were modified to just 0 and 1. Where the value 1 indicates malignant and value 0 indicates benign, turning it into a binary class dataset. After discretization, the class attribute remained to be numeric type. Hence weka.filters.unsupervised.attribute.The numericToNominal filter was used to cover the data type to nominal. Next, weka.filters.unsupervised.attribute.The renameNominalValues filter was used to rename the labels. The attributeIndices field was specified as 10 to select only the class attribute and the value replacements field was set as '2:0,4:1' meaning that replace label 2 with 0 and replace label 4 with 1. The final dataset distribution visualization is shown in Figure 4 below.

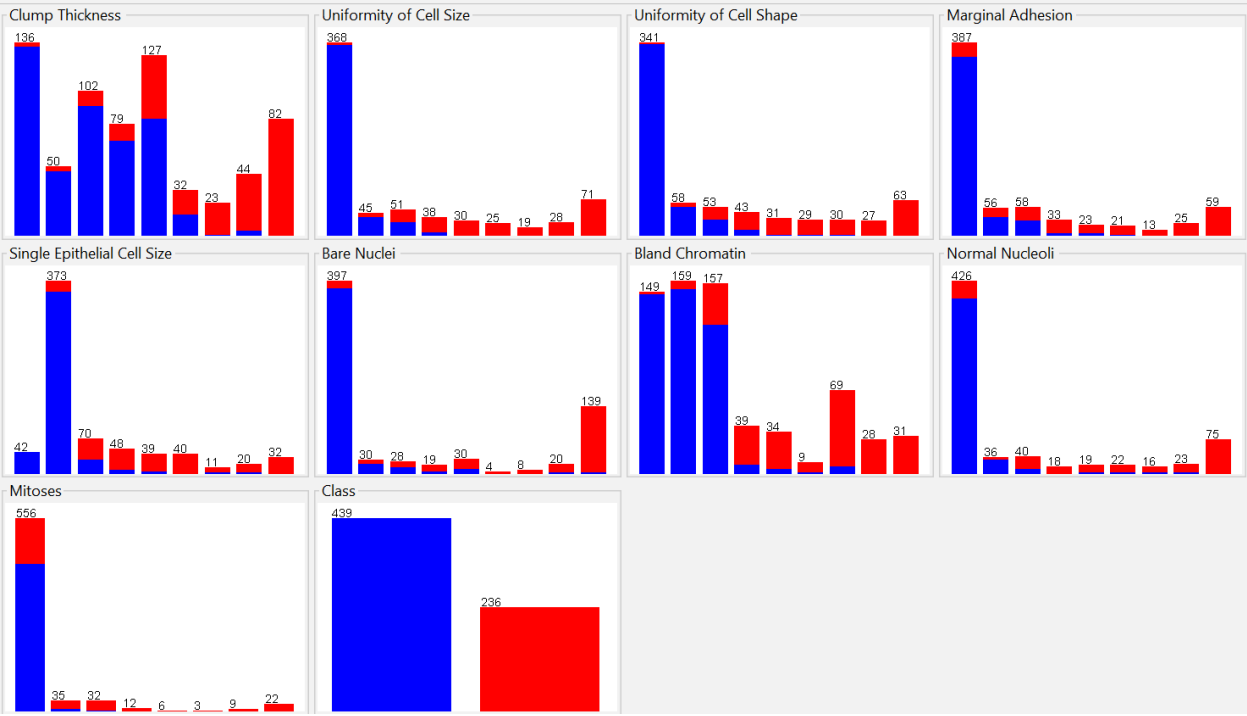

Figure 4: Visualization of the preprocessed dataset attributes

Rule Mining Process

Association Rule Mining (ARM) is a key technique for discovering and extracting useful information from large datasets. It allows users to identify correlations between different objects in databases by generating dependency rules that predict the occurrence of an event based on the

presence of other events. In this study, we apply ARM to understand how various data attributes related to cell nuclei characteristics in breast cancer differentiate between benign (label = 0) and malignant tumors (label = 1).

Association rules are considered significant if they meet user-defined thresholds for minimum support and minimum confidence. Support (supp(X)) measures the proportion of transactions in the dataset that contain a particular item set X, while confidence indicates the certainty of the rule, reflecting the likelihood that the consequent of the rule occurs given the antecedent. For example, an association rule might appear in 47% of transactions (support) and correctly predict the value of the 9th input parameter in 99% of cases where the 1st and 6th parameters are both 1 (confidence).

$$Rule\ X \Longrightarrow Y$$

$$Support = \frac{Frequency\ (X,Y)}{N}$$

$$Confidence = \frac{Frequency\ (X,Y)}{Frequency(X)}$$

In this study, the Apriori algorithm will be used for Association Rule Mining, consisting of two main steps. First, the algorithm applies a minimum support threshold to identify all frequent item sets in the dataset. Next, it uses these itemsets along with the minimum confidence threshold to generate and refine rules. The process involves generating item sets, calculating their frequency, retaining those that exceed the support threshold, and then creating rules from these itemsets while filtering out those with insufficient confidence.

Parameter setting and rule mining

Based on the literature, we initially set a high minimum support value and gradually decrease it until we identified a sufficient number of rules. The minimum support is not fixed; the Apriori algorithm begins with an upper bound for minimum support (default is 1.0, or 100%) and iteratively reduces it by a specified delta (default is 0.05, or 5%). The process continues until the lower bound minimum support is reached or the desired number of rules is generated. The minimum confidence represents the reliability of the derived rules, so we start with a threshold

greater than 60%, as rules with lower confidence are generally considered unreliable. In general, higher confidence values indicate more reliable rules.

After loading the preprocessed dataset into Weka Explorer, we used the Associate tab for rule mining. Given that the dataset is not very large, we started with a minimum support of 30% and a minimum confidence of 80% to identify meaningful and relevant rules. The minimum confidence was set by selecting the 'metricType' as Confidence and 'minMetric' as 0.8. We limited the number of rules to 4. The 'lowerBoundMinSupport' was set to its default value of 0.1, and the 'upperBoundMinSupport' was set to 0.3.

The 'car' field was set to True to mine class association rules that distinguish between benign and malignant tumors, with the 'classIndex' field set to 10 to specify the correct attribute for classification output. Verbose mode was enabled to ensure that the algorithm provided detailed output during its execution. Figure 5 below shows the initial parameter settings for the analysis.

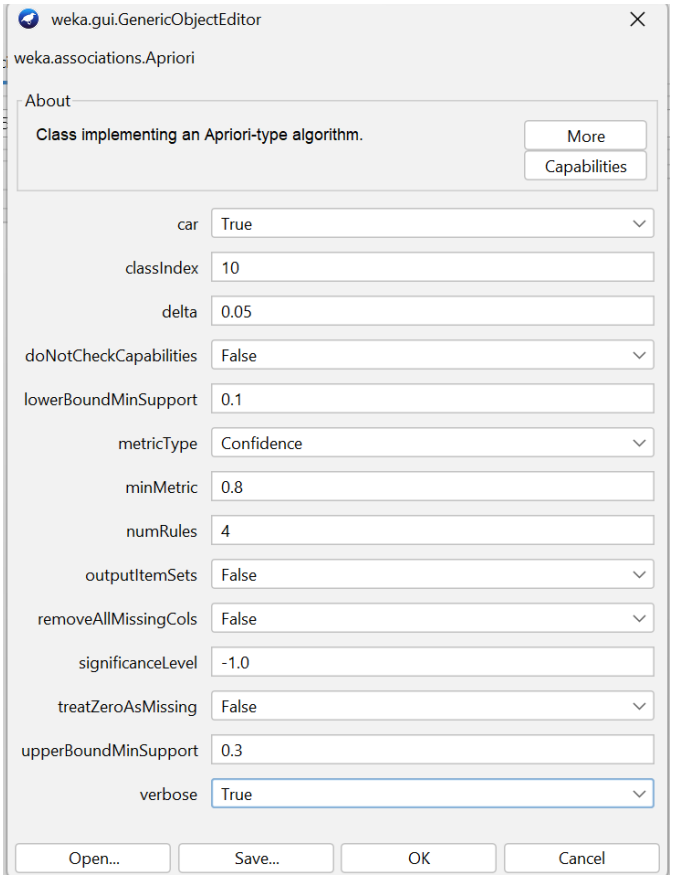

Figure 5: Initial parameter setting for Apriori Algorithm in ARM

After running the algorithm, results appeared on the console within 3-4 seconds. The generated rules had a confidence of 1, indicating high reliability. The algorithm terminated after 16 cycles at a minimum support of 0.2 (135 instances). However, all the best rules identified classified tumors as class 0 (benign). Notably, none of the rules identified malignant tumors (class 1) given the specified support and confidence values, likely due to the low proportion of malignant instances compared to benign ones (approximately 1:2).

Since the initial iteration did not produce any rules distinguishing malignant cancer, the 'lowerBoundMinSupport' was adjusted to 0.05 and the 'upperBoundMinSupport' to 0.1 (10%), while maintaining the confidence at its previous level. The algorithm was run again, but the results still only classified tumors as class 0 (benign) after 18 cycles. Consequently, the parameters were further optimized and a new iteration was conducted.

In the third iteration, the confidence was maintained at 80%, with the 'upperBoundMinSupport' set to 5% and the 'lowerBoundMinSupport' set to 0.005. These settings successfully generated 4 rules with 100% confidence: two rules for class 0 (benign) and two for class 1 (malignant). The algorithm stopped at a minimum support value of 6% and a minimum confidence of 80%, producing valid and reliable rules. This iteration involved 19 cycles. After finding the optimal parameters, the rule number was increased from 4 to 10 to generate more rules and explore additional possibilities. Increasing the 'upperBoundMinSupport' to 0.07 resulted in the absence of rules for class 1 (malignant). Thus, the final optimized parameters were set to a minimum support of 6% and a confidence value of 80%. Further decreasing the confidence did not yield changes, as all generated rules already had 100% confidence. Each run of the algorithm is completed within a few seconds.

Resulting rules

After all the iterations for parameter optimization (support and confidence), four rules were selected out of the 10 rules generated. To select the best four rules out of the ten obtained, it is important to consider the diversity of the conditions (antecedents) leading to the same class, as well as the interpretability and uniqueness of each rule. Since all the rules have perfect confidence (1), the variety in the attributes used was prioritized. The selected rules and their description are as follows:

Rule 1: Uniformity of Cell Size=(8.2-inf) AND Normal Nucleoli=(8.2-inf) ==> Class=1 (confidence = 100%)

This rule indicates that when both the uniformity of cell size and normal nucleoli are in the highest range, the tumor is classified as malignant (Class=1). Uniformity of cell size and the appearance of nucleoli are critical indicators of abnormal, potentially cancerous cell growth. This rule is selected because it highlights key characteristics of malignant tumors, involving two significant features that strongly correlate with cancer. The combination of high values for both attributes makes it a powerful predictor of malignancy.

Rule 2: Marginal Adhesion=(-inf-1.9] AND Single Epithelial Cell Size=(-inf-1.9] ==> Class=0 (confidence = 100%).

This rule suggests that when both marginal adhesion and single epithelial cell size are in the lowest range, the tumor is classified as benign (Class=0). Low marginal adhesion indicates that cells are not sticking together strongly, which is a characteristic of benign tumors, and small epithelial cells are less likely to be malignant. This rule is chosen for its clear representation of benign tumors. It uses two distinctive features that, when both are low, confidently suggest a non-cancerous condition, providing a strong contrast to the rules indicating malignancy.

Rule 3: Normal Nucleoli=(8.2-inf) AND Mitoses=(-inf-1.9] ==> Class=1

This rule indicates that when normal nucleoli are highly prominent (in the highest range) but mitoses are minimal (in the lowest range), the tumor is classified as malignant (Class=1). The presence of large nucleoli is often a sign of malignancy, while the low number of mitoses could indicate a less aggressive, but still malignant, tumor. The selected rules include patterns for both benign (Class=0) and malignant (Class=1) classifications, ensuring that both outcomes are well-represented. Each rule is based on clear, interpretable features that are clinically relevant, making these rules useful for practical application in diagnostic settings.

Rule 4 : Clump Thickness=(-inf-1.9] AND Uniformity of Cell Shape=(-inf-1.9] AND Bland Chromatin=(2.8-3.7] ==> Class=0 (confidence = 100%)

This rule states that if clump thickness and uniformity of cell shape are both in the lowest range, and bland chromatin is in a moderate range, the tumor is likely benign (Class=0). Thin clumps and uniform cell shape are typically associated with benign growths, while moderately bland chromatin

indicates less variability in the appearance of the cell's nucleus. This rule is included because it captures a specific combination of attributes that collectively suggest a benign condition. It adds variety to the selected rules by incorporating three attributes and representing a different aspect of benign tumors. This rule was selected because it highlights the presence of large nucleoli as a strong indicator of malignancy, despite the lower rate of cell division (mitoses). It complements the first rule by focusing on a different aspect of cancerous cells.

Recommendations

Based on the rules discovered, it is recommended to incorporate the discovered rules into the diagnostic process for breast cancer. These rules can be used by clients in healthcare as decision support tools, helping clinicians to more accurately and quickly determine whether a tumor is likely to be benign or malignant based on specific cell features. This can reduce reliance on subjective assessments by experienced doctors alone, potentially lowering the rate of diagnostic errors and ensuring more consistent, data-driven decisions. Further, the rules found can be used to provide training to healthcare practitioners on the significance of the features identified by the rules (e.g., clump thickness, uniformity of cell size, and nucleoli characteristics) and how these should influence their clinical decisions. Education will ensure that healthcare providers understand and trust the data-driven insights, leading to better-informed decision-making in clinical practice.

Clients should test the discovered rules across different populations or datasets to ensure their robustness and generalizability. This could involve cross-validation with other datasets or applying the rules in different clinical settings. Validation will confirm whether the rules are universally applicable or if they need adjustments to account for variability in tumor characteristics across different populations.

Further, the clients can prioritize the high-risk cases such as tumors with high Uniformity of Cell Size and high Normal Nucleoli or High values of Normal Nuclei with low values of Mitoses for immediate further testing and intervention. Cases matching the malignant (Class=1) patterns should be flagged for prompt attention. Early identification and treatment of potentially malignant tumors can improve patient outcomes by enabling timely interventions. The rules can be

incorporated into screening programs to more accurately stratify patients into risk categories. Patients with benign (Class=0) indicators could be monitored with less frequent or less invasive follow-ups, while those with malignant indicators could be scheduled for more rigorous testing. This targeted approach can optimize resource allocation, focusing intensive diagnostic efforts on those who need it most, while reducing unnecessary procedures for low-risk patients.

The client could consider developing or refining predictive models for breast cancer diagnosis using these rules as part of the model's knowledge base. These models could be implemented in software tools used by radiologists and pathologists. Incorporating these rules into predictive analytics tools can enhance the accuracy and reliability of automated or semi-automated diagnostic systems. This information from the discovered rules could be used as a basis for further research to explore the underlying biological mechanisms that cause these specific cell features to correlate with benign or malignant tumors. This could involve genetic studies or investigations into tumor biology. Understanding the biological rationale behind these rules could lead to new insights into breast cancer development, potentially informing new treatment strategies or biomarkers for early detection.

References

Ed-daoudy, A., & Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1), 34.

Kabir, M. F., Ludwig, S. A., & Abdullah, A. S. (2018). Rule discovery from breast cancer risk factors using association rule mining. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2433-2441). IEEE.