INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 04-03-2023

Accepted: 14-08-2023

Published: 27-09-2023

Citation: Alagukumar S, Kathirvalavakumar T (2023) Modified Associative Classification Model for Microarray Gene Expression Data using Maximal Frequent Itemsets and Probability Distribution. Indian Journal of Science and Technology 16(36): 2988-3001. https://doi.org/10.17485/IJST/v16i36.463

* **Corresponding author**.

alagukumarmca@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Modified Associative Classification Model for Microarray Gene Expression Data using Maximal Frequent Itemsets and Probability Distribution

S Alagukumar[1,2]*, **T Kathirvalavakumar**[3]

1 Research Scholar, Research centre in Computer Science, V. H. N. Senthikumara Nadar College, Virudhunagar, Madurai Kamaraj University, Madurai, Tamil Nadu, India
2 Assistant Professor, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India
3 Associate Professor, Research Centre in Computer Science, V. H. N. Senthikumara Nadar College, Virudhunagar, Madurai Kamaraj University, Madurai, Tamil Nadu, India

Abstract

Objective: To make study on generating a less number of class association rules and predicting the class. **Methods:** Modified associative classification model (MACM) is proposed here for diagnosing cancer from microarray gene expression data using maximal frequent itemsets and probability distribution. The proposed system performs supervised discretization, maximal frequent itemset generation from 80% of the data and prediction processes on the 20% of the dataset. The frequent items set are generated using the minimum support as 20%, 40% and 80% and the minimum confidence as 80%. Binary class data sets and multi class data sets are used to evaluate the constructed model and compared with the classical associative classification algorithms. The model performance is evaluated with type of frequent itemset, number of class association rules generated, accuracy and time taken during training the model. The experiment uses the two colorectal cancer datasets, one lung cancer dataset and one multi label cancer datasets. **Findings:** The maximal frequent itemset generates the class association rules quickly with lesser number and leads to consume lesser memory space. The performance of the proposed method provides 100%, classification accuracy for the colon cancer datasets GSE15781 and GSE25070 and 99.17% for the colon cancer data set GSE87211. 94% classification accuracy is obtained for the lung cancer dataset GSE43580 when used maximal frequent itemset types. **Novelty**: Proposed Modified associative classification model has achieved very high performance in classifying gene expression data. The associative classification model helps to diagnose cancer diseases, pathway analysis and treat the cancer disease.

Keywords: Microarray; Discretization; Maximal frequent itemsets; Association rules; Probability distribution

1 Introduction

Early treatment of cancer increases the possibility of curing and reduces the fatality rate and cancer recurrence[1]. An efficient tool is required to diagnose the patients whether they are affected by cancer or not and distinguish different types of cancer. Data mining provides data analysis to uncover interesting knowledge for understanding and diagnosing diseases from microarray gene expression data. However, due to a large number of genes and small samples size of microarray data, the conventional statistical and classification techniques may not be able to deal with it efficiently[2]. Due to huge number of features or genes in the microarray, data may produce redundant results and affect the classification accuracy. To overcome this problem, feature selection, dimension reduction, informative genes identification[3], discretization and rules generation are important in building an associative classification model for gene expression data. Identifying candidate genes for the specific cancers using an informative genes and class associative rules are used to analyze the gene expression with biological information from the gene ontology[4]. The associative classification methods help the healthcare professionals in identifying the cancer risk factorsand diagnose the genes which are cause for diseases. Veroneze et al.[5] have used association rule mining to identify molecular profiles patterns from gene expression data on chronic inflammatory diseases. Luna et al.[6] have reviewed various frequent itemset mining and pointed out that frequent itemset is an essential task for extracting frequently occurring events and patterns in data. However, the computational complexity of the frequent itemsets algorithms is increased exponentially when the data size is increased. Shan and Miao[7] have stated that the class association rule is a special type of association rule suitable for classification problems. Many existing class association rule mining algorithms have inefficiencies when dealing with rules and takes long time to generate the rules. Kenmogne et al.[8] have pointed out that extracting frequent patterns is important in large databases and presented an algorithm for discovering gradual patterns using maximal frequent itemset by reducing search space and the computational time. Alagukumar et al.[9] have used frequent itemsets to generate class association rules for the gene expression data and pointed out that more rules can be minimized by using closed frequent itemsets or maximal frequent itemsets instead of normal frequent itemset.

Sen et al.[10] have analyzed the dengue fever gene expression data using associative classification method and pointed out that the process of mining all frequent itemsets is space and time-consuming when frequent itemsets along with Apriori or FP-Growth algorithm is used. It is challenging to generate effective rules from the correlated genes. Abdo et al.[11] have pointed out that frequent pattern mining is a significant research topic in the medical field and proposed a compressed maximal frequent pattern for Corona virus disease (COVID-19) dataset. Existing associative classification approaches are generating large number of rules from the frequent itemset, especially for dense dataset and occupies more memory space and consumes more execution time. Generating frequent itemsets consume more memory and time. It leads to get research attention in mining frequent itemsets. Also, existing methods are using discrete data to generate class association rules. In this paper, a modified associative classification method is proposed for diagnosing diseases using maximal frequent itemset and probability distribution to generate less number of class association rules. It leads to consume less time with minimum memory space. The proposed method uses the discretized intervals for generating class association rules and is used for identifying the expressed and unexpressed genes among class association rules. The paper is organized as follows. Proposed methodology is presented in Section 2. Section 3 discusses the experimental results. Conclusions are given in section 4.

2 Methodology

2.1 Data Set

The National Center for Biotechnology and Information (NCBI) Gene Expression Omnibus (GEO) database is a public functional genomics database with high-throughput gene expression data, chips, and microarrays. GSE15781[12], GSE87211[13], GSE25070[14] and GSE43580[15] was downloaded from GEO. The GSE15781 dataset consists of colorectal cancer patients includes 13 cancer tissue and 10 normal tissues. The dataset GSE87211 contains 203 colorectal cancer samples and 160 control samples. The GSE25070 dataset contains the 26 tumor samples and 26 normal samples. In this research, the proposed method tested with the multiclass dataset. The GSE43580 gene expression profiles datasets contain 77 lung adeno carcinoma and 73 lung squamous cell carcinoma samples from Gene Expression Omnibus (GEO) under accession number GSE43580. NCI60 is a data set of gene expression profiles of 60 National Cancer Institute (NCI) cell lines. The data set[16] consists of 7129 genes and 60 samples. The 60 human tumor cell lines are divided into eight different cancer classes such as, eight is breast cancer, six CNS cancer, seven colon cancer, six leukemia cancer, eight melanoma cancer, nine non-small-cell lung carcinoma cancer, six ovarian cancers, eight renal cancer tumors and two prostate cancers.

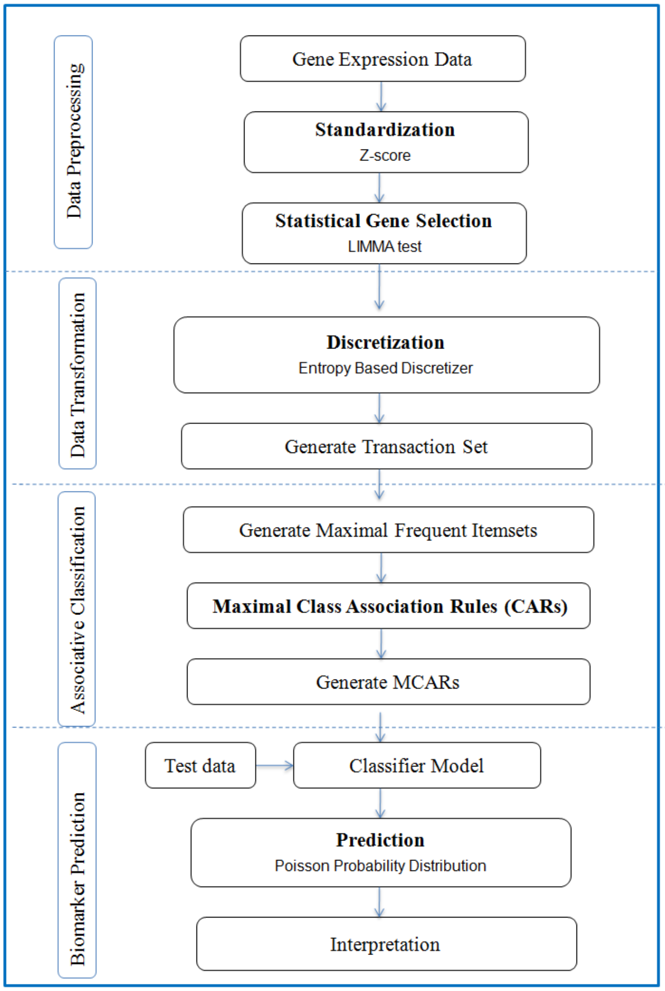

Fig 1. An Associative Classification Model

2.2 Modified Associative Classification Model

The objective of this work is to present an associative classification method for gene expression data to enhance the performance, time and space in the high dimensional data set. In this section, the new proposed method, namely Modified Associative Classification Model (MACM) for microarray gene expression data classification. The proposed model has four phases namely data preprocessing, data transformation, associative classification, and biomarker prediction**. Figure 1**depicts the phases.

2.3 Data Preprocessing

The gene expression data are standardized using the z-score standardization. The standardization brings all genes within a range. After that, informative genes are identified using the LIMMA package which is used to reduce the dimensionality of the dataset.

2.3.1 Data Standardization

Normalization and standardization methods are applied to remove certain systematic biases that are inherent on the data. Before analyzing the data, the gene expression data must be normalized to avoid large variation in the gene expressions and to avoid errors during data processing(17). The Z-score(17) is used for standardizing data and makes significant changes in the gene expression between different samples and conditions. Z score data standardization formula is given as follows:

$$Z = \frac{D - \mu}{\sigma} \tag{1}$$

where, D is the data to be normalized, μ is the arithmetic mean and σ is the standard deviation of that data and Z is the standardized variable with mean 0 and variance 1. This method is used to normalize the gene expression data.

2.3.2 Statistical Gene Selection

Gene selection is a main task for microarray data classification to find the differentially expressed genes and to reduce dimensionality by removing irrelevant and noisy data[(18)]. The Linear Models for Microarray Data (LIMMA) package is R-based open-source software in statistical genomics[(19)]. LIMMA package uses linear models to preprocess and analyze the microarray experiments[(19)]. The LIMMA model requires design matrix and contrast matrix. The first step is to fit a linear model using $E\left(Y_i\right) = X\alpha_j$, where Y_i contains the expression data for the gene j, X is the design matrix and α_j is the vector of coefficient, The coefficients component of the fitted model is produced by linear model. Define $\beta_j = C^T\alpha_j$, where C is the contrast matrix. The linear model for gene j has residual variance σ_j^2 with sample value s_j^2 and degree of freedom f_j. The limma uses moderated t test. The moderated t-statistic is used for significance analysis, and is computed for each gene and for each contrast. The empirical Bayes method assumes an inverse Chi-square prior value of the σ_j^2 with the mean s_0^2, degrees of freedom f_0 and $\widetilde{S}_j^2$ the posterior values for the residual variances which are calculated using Equation (2).

$$\widetilde{S}_j^2 = \frac{f_0 s_0^2 + f_j s_j^2}{f_0 + f_j} \tag{2}$$

The moderated t-statistic for the k$^{\text{th}}$ contrast for the gene j is calculated using Equation (3).

$$t_{jk} = \frac{\beta_{jk}}{U_{jk}\widetilde{S}_j} \tag{3}$$

where U_{jk} is the unscaled standard deviation. The moderated t test follows t-distribution on $f_0 + f_j$ degree of freedom if β_{jk} is equal to 1. The output of empirical Bayes method contains t_{jk} and corresponding p-value.

2.4 Data Transformation

Data transformation[(20)] is used to integrate various types of data and to apply association rule mining successfully in the rule generation phase. Data discretization is a data transformation method, where the gene expression data are transformed from continuous data into nominal data. There are several discretization techniques such as equal width binning, cluster-based discretization, equal depth discretization, class attribute contingency discretization and entropy-based discretization etc. The entropy-based discretization is a supervised approach that discretizes attributes using the class information.

The discretization process[(21)] follows four steps, such as sorting the continuous values, calculating cut points for splitting intervals or merging intervals, based on some condition or criterion, and finally stopping at some point based on the splitting or merging intervals. Gene expression data sets are continuous variables and measured by the interval. The process of partitioning continuous variables into categories is known as discretization. The discretization techniques are located along two dimensions such as supervised versus unsupervised and local versus global. In this work, the entropy-based discretization is used.

2.4.1 Entropy-based Discretization

Entropy-based discretization is a supervised technique, which uses the class information to transform the data into nominal data. The entropy-based discretization process is explained in the following algorithm 1.

```
Algorithm 1: Entropy Based Discretization
Input: Gene Expression with samples and filtered genes
Output: Discretized Gene Expression
Process:
Begin
Step1: Read the statistically and significantly expressed genes
Step2: Sort the gene expression values
Step3: Calculate the entropy $H(X)$ for gene expression data using Equation (4).
```

$$H\left(X\right) = -\sum_{i=0}^{n} p_i * log_2\left(p_i\right) \tag{4}$$

Step4: Search a suitable cut point with the lowest entropy

Step5: Split the range of continuous gene expression values according to cut point is calculated using Equation (5).

$$Info(S,T) = -p_l \sum_{j=1}^{m} p_{j,\ l} log_2 p_{j,\ l} - p_r \sum_{j=1}^{m} p_{j,\ r} log_2 p_{j,\ r} \tag{5}$$

Step6: Repeat steps 4-5 until satisfy the stopping criteria and discretize all the continuous values
End.

2.5 Associative Classification

The filtered and discretized microarray data set is transformed into a transaction set. A microarray transaction table is built based on the class labels before rule mining is applied. The number of transactions in a gene expression dataset corresponds to its number of samples. The number of transactions of biological information is the total number of discrete data acquired. These sets of transactions are passed to the maximal class rule generation phase for model building. Association rule[(22)] and classification are combined into associative classification. Generating a frequent itemset is the one of the key processes in the association analysis to identify the interesting set of genes. Mining frequent itemsets is essential for discovering class association rules. Many of the frequent itemset generation algorithms follow Apriori[(23)], which uses a bottom-up and breadth-first search approach. Generating long frequent patterns in dense data is computationally infeasible. A solution to this problem is to mine only the maximal frequent itemsets[(24)]. The maximal frequent itemset is the frequent itemset for which none of its immediate supersets are frequent, and the maximal pattern set is less than all frequent patterns. Maximum frequent itemset sentences helps to understand long patterns in gene expression data. The process of Class Association Rules follows two steps (1) generate the maximal frequent itemsets and (2) build a classifier from the class association rules. The procedure of the maximal class association rules is explained in the algorithm 2.

2.5.1 Maximal Frequent Itemset

Given a set of items I= {i1, i2, i3 … in} and a set of transaction T = {t1, t2, t3 … tm}, a subset of I is called a frequent, if support(S) ≥ minimum support, where minimum support is a user defined threshold. The maximal frequent itemset[(24)] is smaller than the frequent closed itemset and frequent itemset.

The method to generate maximal frequent itemsets follows a depth first search approach. The frequent itemset is maximal if it is frequent but none of its proper supersets is frequent.

2.5.2 Maximal Class-Association Rules

The maximal classifier model is built from the maximal frequent itemset. A class association rule is the form of $A \to C$, For a rule $A \to C$, A is called an antecedent of the rule, the antecedent of the rule must contains the gene itemsets and C is called a consequent of the rule, the consequent of the rule must contains class labels. The rules are filtered using the confidence threshold given by the user. In this work 20%, 40% and 80 % of support and confidence are used to obtain the maximal class association rules.

The confidence of rule $A \to C$ is computed by calculating the co-occurrence of transactions A and C within the dataset in ratio to transactions containing only A. The confidence of the class association rule is calculated using Equation (6). Finally, the class association rules are regenerated using the Equation (7). In this equation, r_n^c represents class association rules for each class and $MaxR^c$ represents total number of maximal class association rules.

Algorithm 2: Maximal Class Association Rules
Input: Gene Expression with samples and Gene values with intervals
Output: Class Association rules and Classifier Model
Process:

```
Begin
Step 1: Read discretized dataset for each class$c$
Step 2: Transform the dataset into transaction set
Step 3: For each class compute the maximal frequent itemsets
Step 4: Generate a set of rules that have confidence above the minimum confidence threshold from maximal frequent items
```

$$Confidence\ (A \to C) = \frac{s(A \cup C)}{s(A)} \tag{6}$$

Step5: Make a classifier model from these Class Association Rules

$$MaxR^c = (r_1^c, r_2^c \ldots, r_n^c\} \tag{7}$$

```
Step6: Repeat steps 4-6 until form the maximal class association rules for all classes
    End.
```

2.6 Biomarker Prediction

Prediction in associative classification is one of the important steps to determine the accuracy for the developed model. During prediction a sample is predicted to be a particular class when it satisfies more number of eligible rules of the concerned class otherwise it is declared as the default class, which is the majority class in the dataset. Assigning default classes to a sample can affect classifier accuracy. The challenge is to make use of the generated rules in the model to produce a good accuracy. In this paper, a probability-based prediction method in associative classification is proposed. Poisson probability distribution predicts the probability of occurrence of certain events when how often the event has occurred is known. It gives us the number of occurrences of the event in a fixed interval. The Poisson probability distribution is calculated using the Equation (8).

$$P(x;\mu) = \frac{e^{-\mu} * \mu^{x}}{x!} \tag{8}$$

Where, μ is the expected number of occurrences in the rule,e is the base of the logarithm 2.71828 and x represents the test data. The prediction phase of the associative classification is explained in algorithm 3.

```
Algorithm 3: Prediction using probability distribution
Input: Classification Association Rules (Model) Unknown Sample Data set (Test Data)
Output: Classified gene expression
Process:
Begin
Step1: Read Unknown Sample
Step2: Read the Class Association Rules and evaluates how many rules are satisfied in each class
Step3: Assign the test data to that class, whose rules are satisfied maximally using Poisson probability formula
Step4: Repeat steps 2-3 until classify all the test data.
End.
```

3 Results and Discussion

The data of the microarray are presented in the gene expression matrix. Experiments for the proposed method are carried out by the R statistical programming language. Table 1 represents the overview of microarray cancer data sets. Table 2 represents the sample gene expression data matrix. The proposed method can be captured in four phases.

First, in the data preprocessing phase, the raw microarray data were normalized using Z-score normalization and candidate gene features were selected from the normalized data using the LIMMA test. Selected candidate gene features can achieve the highest classification accuracy with the fewest number of genes. Table 3 shows the candidate genes selected using the LIMMA test. There are 10 significant genes extracted using p-value < 0.001. The selected significant gene features are highly correlated with colon cancer. The selected gene features are discretized in the data transformation phase using the entropy-based discretization method. Table 4 shows the candidate gene features discretized by gene intervals. The significant gene expressions are discretized into several intervals using entropy-based discretization algorithm. Table 5 shows that the discrete values for selected candidate genes. The discretized data are converted into the transactional dataset, which are used to generate the frequent item set and class association rules. The frequent items set are generated using the minimum support as 20%, 40% and 80% and the minimum confidence as 80%. Finally, the class association rules are generated, and are used to classify the test dataset using probability distributions.

Table 1. Overview of the microarray datasets

Dataset	GSE ID	No. of samples	No. of tumor sample	No. of normal samples
Colon Cancer	GSE15781	23	13	10
Colon Cancer	GSE87211	363	203	106
Colon Cancer	GSE25070	52	26	26
Lung Cancer	GSE43580	150	73	77

Table 2. Gene Expression Data Set

Sample	class	PCSK1N	C9orf100	LGALS4	….	SPIB	AKR1B10	INSL5	HSD17B2	SLC4A4	KRT20
GSM396309	1	-0.97	-1.08	-0.55	….	-0.84	-0.86	-0.89	-1	-0.93	-0.42
GSM396310	1	-0.63	-0.73	-0.96	….	-0.47	-1.11	-0.91	-1.15	-0.96	-0.89
GSM396311	1	-1.1	-0.8	-0.92	….	-0.95	-0.72	-0.91	-1.18	-0.9	-0.74
GSM396312	1	-0.18	-0.64	0.05	….	-0.8	-0.76	-0.06	0.24	0.49	0.12
GSM396313	1	-1.23	-0.97	-1.01	….	-0.92	-1.22	-0.91	-0.96	-0.93	-1.15
GSM396314	1	0.39	-0.16	0.33	….	0.45	0.07	0.39	0.32	-0.04	0.49
GSM396315	1	-0.38	-0.83	-0.96	….	-0.91	-0.97	-0.85	-0.86	-0.75	-1.06

Table 3. Selected Significant Genes using LIMMA Test

Sample	class	PCSK1N	C9orf100	LGALS4	PDGFD	SPIB	AKR1B10	INSL5	HSD17B2	SLC4A4	KRT20
GSM396309	1	-0.97	-1.08	-0.55	-1.36	-0.84	-0.86	-0.89	-1	-0.93	-0.42
GSM396310	1	-0.63	-0.73	-0.96	-1.39	-0.47	-1.11	-0.91	-1.15	-0.96	-0.89
GSM396311	1	-1.1	-0.8	-0.92	-1.39	-0.95	-0.72	-0.91	-1.18	-0.9	-0.74
GSM396312	1	-0.18	-0.64	0.05	0.16	-0.8	-0.76	-0.06	0.24	0.49	0.12
GSM396313	1	-1.23	-0.97	-1.01	-0.67	-0.92	-1.22	-0.91	-0.96	-0.93	-1.15
GSM396314	1	0.39	-0.16	0.33	0.03	0.45	0.07	0.39	0.32	-0.04	0.49
GSM396315	1	-0.38	-0.83	-0.96	-0.54	-0.91	-0.97	-0.85	-0.86	-0.75	-1.06
GSM396316	1	-0.77	-0.36	-1.14	-0.01	-0.8	-0.82	-0.84	-0.83	-0.86	-0.91
GSM396317	1	-1.14	-0.7	-0.93	-1.25	-1.04	-1.15	-0.91	-0.8	-0.89	-1.12
GSM396318	1	-1.3	-0.41	-1.3	-1.36	-0.97	-1.22	-0.92	-0.96	-0.94	-1.13
GSM396319	1	-0.95	-0.98	-0.79	-0.05	-0.83	-0.07	-0.9	-0.63	-0.84	-0.78
GSM396320	1	-0.44	-0.97	-0.6	-0.5	-0.57	-0.07	-0.7	-0.51	-0.81	-0.65
GSM396321	1	-0.84	-0.68	-0.52	-0.95	-0.61	-0.35	-0.78	-0.86	-0.74	-0.85
GSM396322	0	0.87	0.82	1.43	0.74	1.31	1.81	1.32	1.62	2.15	0.93
GSM396323	0	1.47	1.72	-0.07	0.32	1.85	-0.27	1.57	0.15	0.4	1.03
GSM396324	0	0.5	0.18	0.97	1.38	1.54	1.11	-0.11	1.76	0.96	0.98
GSM396325	0	0.59	-0.09	0.62	0.4	0.53	0.5	1.65	0.64	0.23	2.52
GSM396326	0	0.77	2.55	1.82	1.3	0.14	2.04	2.12	1.88	1.72	1.4
GSM396327	0	0.37	1.16	0.61	1.51	-0.04	0.49	0.39	0.49	0.51	0.12
GSM396328	0	0.58	0.87	1.4	1.04	0.46	0.32	0.46	0.09	0.9	0.32
GSM396329	0	2.09	1	0.58	1.47	0.95	1.11	1.23	1.43	1.06	0.34
GSM396330	0	0.58	1.08	0.1	0.07	0.41	1.09	0.06	0.2	-0.37	0.47
GSM396331	0	1.71	0.02	1.85	1.04	2.12	1.06	0.48	0.91	1.53	0.98

Table 4. Discretized Data using Entropy

Sample	class	PCSK1N	C9orf100	LGALS4	PDGFD	SPIB	AKR1B10	INSL5	HSD17B2	SLC4A4	KRT20

Continued on next page

Table 4 continued

GSM396309	1	0, 0.095	-Inf,- 0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396310	1	0, 0.095	-Inf,- 0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396311	1	0, 0.095	-Inf,- 0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396312	1	0, 0.095	-Inf -0.125	-0.010 0.075	0.115 0.240	-Inf -0.255	-Inf -0.31	-0.085 0.000	0.220 0.405	0.445 0.500	-0.15 0.22
GSM396313	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396314	1	0.380, 0.445	-Inf -0.125	0.215 0.455	-Inf 0.05	0.430 0.455	-0.170 0.195	0.225 0.425	0.220 0.405	-0.205 0.095	0.48 0.71
GSM396315	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396316	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396317	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396318	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396319	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-0.170 0.195	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396320	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-0.170 0.195	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396321	1	0, 0.095	-Inf -0.125	-Inf -0.295	-Inf 0.05	-Inf -0.255	-Inf -0.31	-Inf -0.405	-Inf -0.21	-Inf -0.555	-Inf -0.15
GSM396322	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	0.425 Inf	0.405 Inf	0.5 Inf	0.71 Inf
GSM396323	0	0.445 , Inf	-0.125 Inf	-0.305	0.24 Inf	0.455 Inf	-0.48	0.425 Inf	-0.21 0.22	0.095 0.445	0.71 Inf
GSM396324	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	-0.49	0.405 Inf	0.5 Inf	0.71 Inf
GSM396325	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	0.425 Inf	0.405 Inf	0.095 0.445	0.71 Inf
GSM396326	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	-0.255 0.430	0.195 Inf	0.425 Inf	0.405 Inf	0.5 Inf	0.71,Inf
GSM396327	0	0.095, 0.380	-0.125 Inf	0.455 Inf	0.24 Inf	-0.255 0.430	0.195 Inf	0.225 0.425	0.405 Inf	0.5 Inf	-0.15 0.22
GSM396328	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	0.425 Inf	-0.21 0.22	0.5 Inf	0.22 0.48
GSM396329	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	0.425 Inf	0.405 Inf	0.5 Inf	0.22 0.48
GSM396330	0	0.445 , Inf	-0.125 Inf	0.075 0.215	0.050 0.115	-0.255 0.430	0.195 Inf	0.000 0.225	-0.21 0.22	-0.76	0.22 0.48
GSM396331	0	0.445 , Inf	-0.125 Inf	0.455 Inf	0.24 Inf	0.455 Inf	0.195 Inf	0.425 Inf	0.405 Inf	0.5 Inf	0.71 Inf

Table 5. Discretized Data

Samples	class	PCSK1N	C9orf100	LGALS4	PDGFD	SPIB	AKR1B10	INSL5	HSD17B2	SLC4A4	KRT20
GSM396309	1	1	1	1	1	1	1	1	1	1	1
GSM396310	1	1	1	1	1	1	1	1	1	1	1
GSM396311	1	1	1	1	1	1	1	1	1	1	1
GSM396312	1	1	1	3	3	1	1	3	3	5	2
GSM396313	1	1	1	1	1	1	1	1	1	1	1
GSM396314	1	3	1	5	1	3	3	5	3	3	4

Continued on next page

Table 5 continued

GSM396315	1	1	1	1	1	1	1	1	1	1	1
GSM396316	1	1	1	1	1	1	1	1	1	1	1
GSM396317	1	1	1	1	1	1	1	1	1	1	1
GSM396318	1	1	1	1	1	1	1	1	1	1	1
GSM396319	1	1	1	1	1	1	3	1	1	1	1
GSM396320	1	1	1	1	1	1	3	1	1	1	1
GSM396321	1	1	1	1	1	1	1	1	1	1	1
GSM396322	0	4	2	6	4	4	4	6	4	6	5
GSM396323	0	4	2	2	4	4	2	6	2	4	5
GSM396324	0	4	2	6	4	4	4	2	4	6	5
GSM396325	0	4	2	6	4	4	4	6	4	4	5
GSM396326	0	4	2	6	4	2	4	6	4	6	5
GSM396327	0	2	2	6	4	2	4	5	4	6	2
GSM396328	0	4	2	6	4	4	4	6	2	6	3
GSM396329	0	4	2	6	4	4	4	6	4	6	3
GSM396330	0	4	2	4	2	2	4	4	2	2	3
GSM396331	0	4	2	6	4	4	4	6	4	6	5

In the associative classification phase, the discretized candidate gene sets are applied for associative classification. Before mining the generated rules, the transaction table was partitioned based on the class labels. The MACM algorithm produces the desired number of maximal frequent itemsets with the highest support value from each target class label. These maximal frequent itemsets are applied to generate maximal class association rules using the minimum confidence. Table 6 shows the generated maximal class association rules of the proposed method.

Table 6. Maximal Class Association Rules

Rule No.	LHS	RHS
1.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], SPIB[lower = -0.255, upper = 0.43], AKR1B10[lower = 0.195, upper = Inf]	Class=Control
2.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], AKR1B10[lower = 0.195, upper = Inf], HSD17B2[lower = -0.21, upper = 0.22], KRT20[lower = 0.22, upper = 0.48]	Class=Control
3.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], PDGFD[lower = 0.24, upper = Inf], SPIB[lower = 0.455, upper = Inf], INSL5[lower = 0.425, upper = Inf], HSD17B2[lower = -0.21, upper = 0.22]	Class=Control
4.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], PDGFD[lower = 0.24, upper = Inf], SPIB[lower = 0.455, upper = Inf], INSL5[lower = 0.425, upper = Inf], SLC4A4[lower = 0.095, upper = 0.445], KRT20[lower = 0.71, upper = Inf]	Class=Control
5.	C9orf100[lower = -0.125, upper = Inf], LGALS4[lower = 0.455, upper = Inf], PDGFD[lower = 0.24, upper = Inf], SPIB[lower = -0.255, upper = 0.43], AKR1B10[lower = 0.195, upper = Inf], HSD17B2[lower = 0.405, upper = Inf], SLC4A4[lower = 0.5, upper = Inf]	Class=Control
6.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], LGALS4[lower = 0.455, upper = Inf], PDGFD[lower = 0.24, upper = Inf], SPIB[lower = 0.455, upper = Inf], AKR1B10[lower = 0.195, upper = Inf], INSL5[lower = 0.425, upper = Inf], SLC4A4[lower = 0.5, upper = Inf], KRT20[lower = 0.22, upper = 0.48]	Class=Control
7.	PCSK1N[lower = 0.445, upper = Inf], C9orf100[lower = -0.125, upper = Inf], LGALS4[lower = 0.455, upper = Inf], PDGFD[lower = 0.24, upper = Inf], SPIB[lower = 0.455, upper = Inf], AKR1B10[lower = 0.195, upper = Inf], INSL5[lower = 0.425, upper = Inf], HSD17B2[lower = 0.405, upper = Inf], SLC4A4[lower = 0.5, upper = Inf], KRT20[lower = 0.71, upper = Inf]	Class=Control
8.	C9orf100[lower = -Inf, upper = -0.125],PDGFD[lower = -Inf, upper = 0.05],	Class=Cancer

Continued on next page

Table 6 continued

	AKR1B10[lower = -0.17, upper = 0.195]	
9.	PCSK1N[lower = -Inf, upper = 0.095],C9orf100[lower = -Inf, upper = -0.125], LGALS4[lower = -Inf, upper = -0.295],PDGFD[lower = -Inf, upper = 0.05], SPIB[lower = -Inf, upper = -0.255],AKR1B10[lower = -Inf, upper = -0.31], INSL5[lower = -Inf, upper = -0.405],HSD17B2[lower = -Inf, upper = -0.21], SLC4A4[lower = -Inf, upper = -0.555], SLC4A4 [lower = -Inf, upper = -0.15]	Class=Cancer

It is observed from the Table 6 that the genes PCSK1N, C9orf100, PDGFD, AKR1B10, INSL5, SPIB, HSD17B2, SLC4A4 andSLC4A4are used in determining the disease of being colon cancer. In this study, the proposed maximal associative classification performs better in identifying the genes for the cause of cancer. The biomarker prediction phase uses the Poisson probability distribution to evaluate the test gene samples. Table 7 compares the proposed method MACM, conventional CBA and other frequent itemset types. The performance of the proposed method is measured based on support, confidence, training time, number of class association rules and classification accuracy. Table 7 shows that the maximal frequent itemset generates less number of rules than the frequent itemset and closed frequent itemset for the dataset GSE15781, GSE87211, GSE25070, GSE43580. Table 7 also shows that the Z-score normalization works better than min-max normalization and produces less number of class association rules. It is observed from Table 7 that the proposed maximal frequent itemset method consumes less time than the CBA and other frequent itemset types.

Table 7. Comparative Results of Proposed Method with Other Frequent Types and CBA

Classification	**Normalization/ Standardization**	**Dataset**	**Target Type**	**Support in %**	**Confidence in %**	**Number of CARs**	**Accuracy**	**Error Rate**	**Training Time in Seconds**
MACM	Min-Max	GSE15781	Frequent	20	80	2256	100	0	3.36
MACM	Min-Max	GSE15781	Closed	20	80	102	100	0	0.27
MACM	**Min-Max**	**GSE15781**	**Maximal**	**20**	**80**	**19**	**100**	**0**	**0.12**
MACM	Min-Max	GSE15781	Frequent	40	80	1292	100	0	3.4
MACM	Min-Max	GSE15781	Closed	40	80	66	100	0	0.21
MACM	**Min-Max**	**GSE15781**	**Maximal**	**40**	**80**	**18**	**100**	**0**	**0.14**
MACM	Min-Max	GSE15781	Frequent	80	80	266	100	0	16.78
MACM	Min-Max	GSE15781	Closed	80	80	9	100	0	0.15
MACM	**Min-Max**	**GSE15781**	**Maximal**	**80**	**80**	**6**	**100**	**0**	**25.04**
MACM	Z-Score	GSE15781	Frequent	20	80	2482	100	0	70.8
MACM	Z-Score	GSE15781	Closed	20	80	57	100	0	67.8
MACM	**Z-Score**	**GSE15781**	**Maximal**	**20**	**80**	**9**	**100**	**0**	**0.16**
MACM	Z-Score	GSE15781	Frequent	40	80	1766	100	0	15.33
MACM	Z-Score	GSE15781	Closed	40	80	42	100	0	0.18
MACM	**Z-Score**	**GSE15781**	**Maximal**	**40**	**80**	**9**	**100**	**0**	**0.15**
MACM	Z-Score	GSE15781	Frequent	80	80	532	100	0	4.66
MACM	Z-Score	GSE15781	Closed	80	80	11	100	0	0.15
MACM	Z-Score	GSE15781	Frequent	80	80	4	100	0	0.15
CBA	Z-Score	GSE15781	Frequent	20	80	770	100	0	0.25
CBA	Z-Score	GSE15781	Frequent	40	80	640	100	0	0.25
CBA	Z-Score	GSE15781	Frequent	50	80	19	100	0	0.19
CBA	Z-Score	GSE15781	Frequent	20	80	770	100	0	0.25
CBA	Min Max	GSE15781	Frequent	40	80	547	100	0	0.26
CBA	Min Max	GSE15781	Frequent	50	80	30	100	0	0.21
MACM	Min Max	GSE87211	Frequent	20	80	528	97.79	2.21	3.96
MACM	Min Max	GSE87211	Closed	20	80	342	97.79	2.21	2.57
MACM	**Min Max**	**GSE87211**	**Maximal**	**20**	**80**	**33**	**97.79**	**2.21**	**0.33**
MACM	Min Max	GSE87211	Frequent	40	80	277	96.69	3.31	1.47
MACM	Min Max	GSE87211	Closed	40	80	241	96.69	3.31	1.18
MACM	**Min Max**	**GSE87211**	**Maximal**	**40**	**80**	**18**	**96.69**	**3.31**	**0.28**
MACM	Min Max	GSE87211	Frequent	80	80	4	91.73	8.27	0.17
MACM	Min Max	GSE87211	Closed	80	80	4	91.73	8.27	0.14
MACM	Min Max	GSE87211	Maximal	80	80	2	91.73	8.27	0.14

Continued on next page

Table 7 continued

MACM	Z-Score	GSE87211	Frequent	20	80	541	99.17	0.83	3.82
MACM	Z-Score	GSE87211	Closed	20	80	526	99.17	0.83	3.74
MACM	**Z-Score**	**GSE87211**	**Maximal**	**20**	**80**	**102**	**99.17**	**0.83**	**0.79**
MACM	Z-Score	GSE87211	Frequent	40	80	77	98.62	1.38	0.62
MACM	Z-Score	GSE87211	Closed	40	80	77	98.62	1.38	0.64
MACM	**Z-Score**	**GSE87211**	**Maximal**	**40**	**80**	**29**	**98.62**	**1.38**	**0.31**
MACM	Z-Score	GSE87211	Frequent	80	80	1	55.92	44.08	0.14
MACM	Z-Score	GSE87211	Closed	80	80	1	55.92	44.08	0.14
MACM	Z-Score	GSE87211	Maximal	80	80	1	55.92	44.08	0.14
CBA	Z-Score	GSE87211	Frequent	20	80	770	98.34	1.66	0.44
CBA	Z-Score	GSE87211	Frequent	40	80	615	98.34	1.66	0.3
CBA	Z-Score	GSE87211	Frequent	80	80	115	98.34	1.66	0.3
CBA	Min – Max	GSE87211	Frequent	20	80	770	98.89	1.11	0.37
CBA	Min – Max	GSE87211	Frequent	40	80	320	98.34	1.66	0.42
CBA	Min – Max	GSE87211	Frequent	50	80	14	98.07	1.93	0.27
MACM	Min – Max	GSE25070	Frequent	20	80	2510	100	0	6.02
MACM	Min – Max	GSE25070	Closed	20	80	202	100	0	0.58
MACM	**Min – Max**	**GSE25070**	**Maximal**	**20**	**80**	**13**	**100**	**0**	**0.16**
MACM	Min – Max	GSE25070	Frequent	40	80	1578	100	0	3.85
MACM	Min – Max	GSE25070	Closed	40	80	140	100	0	0.78
MACM	**Min – Max**	**GSE25070**	**Maximal**	**40**	**80**	**23**	**100**	**0**	**0.17**
MACM	Min – Max	GSE25070	Frequent	80	80	1022	100	0	2.68
MACM	Min – Max	GSE25070	Closed	80	80	45	100	0	0.25
MACM	**Min – Max**	**GSE25070**	**Maximal**	**80**	**80**	**11**	**100**	**0**	**0.17**
MACM	Z-Score	GSE25070	Frequent	20	80	2510	100	0	6.36
MACM	Z-Score	GSE25070	Closed	20	80	202	100	0	0.72
MACM	**Z-Score**	**GSE25070**	**Maximal**	**20**	**80**	**13**	**100**	**0**	**0.17**
MACM	Z-Score	GSE25070	Frequent	40	80	1578	100	0	3.79
MACM	Z-Score	GSE25070	Closed	40	80	140	100	0	0.42
MACM	**Z-Score**	**GSE25070**	**Maximal**	**40**	**80**	**23**	**100**	**0**	**0.19**
MACM	Z-Score	GSE25070	Frequent	80	80	1022	100	0	2.54
MACM	Z-Score	GSE25070	Closed	80	80	45	100	0	0.23
MACM	**Z-Score**	**GSE25070**	**Maximal**	**80**	**80**	**11**	**100**	**0**	**0.14**
CBA	Z-Score	GSE25070	Frequent	20	80	770	100	0	0.21
CBA	Z-Score	GSE25070	Frequent	40	80	770	100	0	0.23
CBA	Z-Score	GSE25070	Frequent	50	80	262	100	0	0.23
CBA	Min – Max	GSE25070	Frequent	20	80	770	100	0	0.28
CBA	Min – Max	GSE25070	Frequent	40	80	770	100	0	0.29
CBA	Min – Max	GSE25070	Frequent	50	80	262	100	0	0.24
MACM	Min – Max	GSE43580	Frequent	20	80	121	94	6	0.62
MACM	Min – Max	GSE43580	Closed	20	80	108	94	6	0.48
MACM	**Min – Max**	**GSE43580**	**Maximal**	**20**	**80**	**35**	**94**	**6**	**0.23**
MACM	Min – Max	GSE43580	Frequent	40	80	42	77.33	22.67	0.24
MACM	Min – Max	GSE43580	Closed	40	80	42	77.33	22.67	0.24
MACM	Min – Max	GSE43580	Maximal	40	80	21	77.33	22.67	0.19
MACM	Min – Max	GSE43580	Frequent	80	80	13	48.67	51.33	0.23
MACM	Min – Max	GSE43580	Closed	80	80	13	48.67	51.33	0.17
MACM	Min – Max	GSE43580	Maximal	80	80	7	48.67	51.33	0.15
MACM	Z-Score	GSE43580	Frequent	20	80	282	84	16	1.23
MACM	Z-Score	GSE43580	Closed	20	80	238	84	16	1.13
MACM	Z-Score	GSE43580	Maximal	20	80	40	84	16	0.26
MACM	Z-Score	GSE43580	Frequent	40	80	86	78.67	21.33	0.59
MACM	Z-Score	GSE43580	Closed	40	80	82	78.67	21.33	0.4
MACM	Z-Score	GSE43580	Maximal	40	80	25	78.67	21.33	0.21
MACM	Z-Score	GSE43580	Frequent	80	80	18	48.67	51.33	0.35
MACM	Z-Score	GSE43580	Closed	80	80	18	48.67	51.33	0.2

Continued on next page

Table 7 continued

MACM	Z-Score	GSE43580	Maximal	80	80	12	48.67	51.33	0.15
CBA	Z-Score	GSE43580	Frequent	20	80	769	91.33	8.67	0.49
CBA	Z-Score	GSE43580	Frequent	40	80	343	89.33	10.67	0.31
CBA	Z-Score	GSE43580	Frequent	50	80	1	-	-	0.14
CBA	Min – Max	GSE43580	Frequent	20	80	672	92.67	7.33	0.44
CBA	Min – Max	GSE43580	Frequent	40	80	246	90.67	9.33	0.32
CBA	Min – Max	GSE43580	Frequent	50	80	-	-	-	0.17

Table 8 compares the proposed method MACM and conventional CBA for the multi label datasets. The performance of the proposed method is measured based on support, confidence, training time and number of class association rules generated and classification accuracy.

Table 8. Results on Multi-Classification data set

Classification	Standardization	Target Type	Support in %	Confidence in %	No. of Rules	Accuracy	Error Rate	Time Taken in training data (in seconds)	Time Taken in Test data (in seconds)
MACM	Z-Score	Frequent	10	80	5105	88.89	11.11	105.6	227.4
MACM	Z-Score	Frequent	20	80	5105	33.33	66.67	13.19	27.3
MACM	Z-Score	Frequent	40	80	2046	33.33	66.67	3.66	7.43
MACM	Z-Score	Closed	10	80	94	88.89	11.11	0.65	0.51
MACM	Z-Score	Closed	20	80	94	33.33	66.67	0.49	0.18
MACM	Z-Score	Closed	40	80	17	33.33	66.67	0.37	0.06
MACM	Z-Score	Maximal	10	80	60	88.89	11.11	0.81	0.41
MACM	Z-Score	Maximal	40	80	8	33.33	66.67	0.46	0.051

3.1 Performance Metrics

The proposed MACM model was analyzed using the AUC curve and accuracy. Table 9 depicts the four outcomes of binary classification. The accuracy measure is computed using the Equation (9).

Table 9. Binary classification output description

Classifier Outcome	Descriptions
A	Number of affected tissues that are correctly diagnosed
B	Number of healthy tissues that are wrongly identified as a tissue
C	Number of healthy tissues that are correctly diagnosed
D	Number of affected tissues that are wrongly identified as a healthy

$$\frac{A+B}{A+B+C+D} = \frac{tissues\ diagnosed\ correctly}{total\ tissues} \tag{9}$$

Figure 2 depicts the AUC curve of the proposed methods on the colon and lung cancer data sets. The results obtained are compared with the traditional CBA algorithm. The proposed MACM model provides 100%, classification accuracy for the colon cancer datasets GSE15781 and GSE25070; and 99.17% for the colon cancer data set GSE87211. 94% classification accuracy is obtained for the lung cancer dataset GSE43580. Interpretation of the generated rule is simpler and easier for the bio markers as it is with only significant genes and less in numbers. Another feature of the proposed work is rules are generated with gene expression level; it is not available in the existing methods. This gene expression level is used to find whether the gene is over expressed or under expressed.

The proposed method uses the supervised discretization to generate rules with gene expression intervals. The supervised discretization method uses class information to split data into set of discrete intervals and without loss of information. Compared with frequent itemset and closed frequent itemset types, the class association rule mining algorithm using maximal frequent itemsets. The proposed method does not generate numerous frequent itemsets so that the generation of redundant

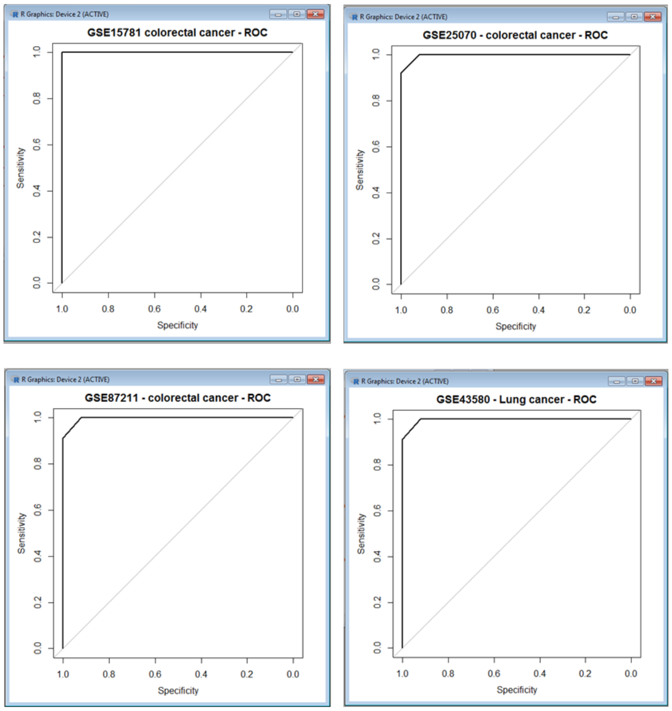

Fig 2. The Area Under the Curve

rules decreases and mining rules becomes faster. The proposed algorithm improves the space and time utilization for the gene expression data and helps to identify the relationship among the gene expression profiles. The proposed method helps to find the functions of the genes and enrichment analysis for group of genes. The Table 10 compares the generated class association rules of the proposed work with the existing methods.

Table 10. Compare the class association rules with the existing methods

Existing Methods	Generated Class Association Rules	Discussions
Ong et al [14]	gene1, gene2, gene3 ,.....genen→ class1gene2, gene3.....genen→class2	Method extracts the association rules with gene name
Yuan, et al [15]	If (TSC2 $\leq$ 124.073) and (GLTP $\geq$ 2042.765) Then →Lung SCC	Extracts the association rules using IF–THEN relationship (e.g., IF gene1 $\geq$ 6.4 AND gene2 $\geq$ 4.8 THEN lung AC)
Proposed MACMMethod	g1[interval], g2[interval], ... gn[interval]→ class1g5[interval], g6[interval],... gn[interval] → class2g1[interval], g2[interval],... gn[interval] → classn **Example :** PCSK1N[0.445-Inf], SPIB[-0.255- 0.43], AKR1B10[0.195-Inf] → normal	Proposed method is applied to extract the association rules with gene expression intervalsIt is used to find correlation among genes expression data profiling and to identify the positive and negative regulators of the genes

4 Conclusion

The proposed method diagnoses diseases from microarray gene expression data using maximal frequent itemsets and probability-based distribution prediction method. Experimental results show that the maximal frequent itemsets quickly generate the rules and consume less memory space for storing the maximal frequent itemsets. Existing methods use only frequent itemsets but the proposed method uses frequent itemsets, closed frequent itemsets and maximal frequent itemsets. The experiments are carried out for the binary class datasets colon cancer and lung cancer; and for the multi class data set National Cancer Institute-60 (NCI-60) cancer cell line gene expression data. The proposed MACM model provides 100% accuracy for the binary class datasets but provides 88.89 % accuracy for the multi-class dataset. In the existing methods, when the generated rules are not matched with the test pattern then they assign the most class frequent in the training data. But as the proposed method uses the probability distribution, it assigns the predicted class for the test pattern based on the probability of the rules covered for a class. Also, the proposed method uses only maximal frequent itemsets which leads to avoid rule pruning. Proposed method works the best only for binary class dataset is its limitation. Future work can be concentrated on multi class data sets and an ensemble soft weighted gene selection-based approach and cancer classification using modified meta-heuristic learning can be proposed for enhancing the current work.

References

1) Fathi H, Alsalman H, Gumaei A, Manhrawy IIM, Hussien AG, El-Kafrawy P. An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data. *Computational Intelligence and Neuroscience*. 2021;2021:1–14. Available from: https://doi.org/10.1155/2021/7231126.
2) Shah SH, Iqbal MJ, Ahmad I, Khan S, Rodrigues JJPC. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Computing and Applications*. 2020;p. 1–12. Available from: https://doi.org/10.1007/s00521-020-05367-8.
3) Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*. 2020;143:1–19. Available from: https://doi.org/10.1016/j.csda.2019.106839.
4) B Çİ, Ö ÜMK, Çolak C. Assessment of COVID-19-Related Genes Through Associative Classification Techniques. *Konuralp Medical Journal*. 2022;14(1):1–8. Available from: https://doi.org/10.18521/ktd.958555.
5) Veroneze R, Corbi SCT, Silva BRD, Rocha CDS, Maurer-Morelli CV, Orrico SRP, et al. Using association rule mining to jointly detect clinical features and differentially expressed genes related to chronic inflammatory diseases. *PLOS ONE*. 2020;15(10):1–22. Available from: https://doi.org/10.1371/journal.pone.0240269.
6) Fournier-Viger LJM, P, S V. Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2019;9(6). Available from: https://doi.org/10.1002/widm.1329.
7) Shan Z, Miao W. COVID-19 patient diagnosis and treatment data mining algorithm based on association rules. *Expert Systems*. 2023;40(4). Available from: https://doi.org/10.1111/exsy.12814.
8) Kenmogne EB, Fotso LCT, Djamegni CT. A novel algorithm for mining maximal frequent gradual patterns. *Engineering Applications of Artificial Intelligence*. 2023;120:105939. Available from: https://doi.org/10.1016/j.engappai.2023.105939.
9) Alagukumar S, Kathirvalavakumar T, Prasath R. Compact Associative Classification for Up and Down Regulated Genes Using Supervised Discretization and Clustering. In: MIKE 2021: Mining Intelligence and Knowledge Exploration, Cham;vol. 13119 of Lecture Notes in Computer Science book series. Springer International Publishing. 2022;p. 33–46. Available from: https://doi.org/10.1007/978-3-031-21517-9_4.
10) Sen D, Paladhi S, Frnda J, Chatterjee S, Banerjee S, Nedoma J. Associative Classifier Coupled With Unsupervised Feature Reduction for Dengue Fever Classification Using Gene Expression Data. *IEEE Access*. 2022;10:88340–88353. Available from: https://doi.org/10.1109/ACCESS.2022.3198937.
11) Abdo AS, Abdul-Kader HM, Salem RK. Enhanced Compressed Maximal Frequent Patterns from COVID-19 Streaming Data. *Studies in Informatics and Control*. 2022;31(1):99–108. Available from: https://doi.org/10.24846/v31i1y202210.
12) Xu H, Ma Y, Zhang J, Gu J, Jing X, Lu S, et al. Identification and Verification of Core Genes in Colorectal Cancer. *BioMed Research International*. 2020;2020:1–13. Available from: https://doi.org/10.1155/2020/8082697.
13) Lv J, Li L. Hub Genes and Key Pathway Identification in Colorectal Cancer Based on Bioinformatic Analysis. *BioMed Research International*. 2019;2019:1–13. Available from: https://doi.org/10.1155/2019/1545680.
14) Ong HF, Mustapha N, Hamdan H, Rosli R, Mustapha A. Informative top-k class associative rule for cancer biomarker discovery on microarray data. *Expert Systems with Applications*. 2020;146:113169. Available from: https://doi.org/10.1016/j.eswa.2019.113169.
15) Yuan F, Lu L, Zou Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2020;1866(8):1–9. Available from: https://doi.org/10.1016/j.bbadis.2020.165822.
16) Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park JMJ, et al. Chemosensitivity prediction by transcriptional profiling. *Proceedings of the National Academy of Sciences*. 2001;98(19):10787–10792. Available from: https://pubmed.ncbi.nlm.nih.gov/11553813/.
17) Maniruzzaman, Rahman JJ, Ahammed B, Abedin M, Suri HS, Biswas MS, et al. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. In: Sinha GR, Suri JS, editors. Cognitive Informatics, Computer Modeling, and Cognitive Science;vol. 1 of Theory, Case Studies, and Applications. Elsevier. 2020;p. 273–317. Available from: https://doi.org/10.1016/B978-0-12-819443-0.00014-3.
18) Abd-Elnaby M, Alfonse M, Roushdy M. Classification of breast cancer using microarray gene expression data: A survey. *Journal of Biomedical Informatics*. 2021;117:1–9. Available from: https://doi.org/10.1016/j.jbi.2021.103764.
19) Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*. 2004;3. Available from: https://doi.org/10.2202/1544-6115.1027.
20) de Sá CR, Soares C, Knobbe A, A. Entropy-based discretization methods for ranking data. *Information Sciences*. 2016;329:921–936. Available from: https://doi.org/10.1016/j.ins.2015.04.022.
21) Garcia S, Luengo J, Sáez JA, López V, Herrera F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2013;25(4):734–750. Available from: https://www.computer.org/csdl/journal/tk/2013/04/ttk2013040734/13rRUwInvJH.
22) Badhon B, Kabir MMJ, Xu S, Kabir M. A survey on association rule mining based on evolutionary algorithms. *International Journal of Computers and Applications*. 2021;43(8):775–785. Available from: https://doi.org/10.1080/1206212X.2019.1612993.
23) Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*. 1993;22(2):207–216. Available from: https://doi.org/10.1145/170036.170072.
24) Burdick D, Calimlim M, Flannick J, Gehrke J, Yiu T. MAFIA: a maximal frequent itemset algorithm. *IEEE Transactions on Knowledge and Data Engineering*. 2005;17(11):1490–1504. Available from: https://doi.org/10.1109/TKDE.2005.183.