INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Impacts of Mutual Information and Discretization Methods in Associative Classifications

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
3 Associate Professor, Research Centre in Computer Science, V. H. N. Senthikumara Nadar College, Virudhunagar, Madurai Kamaraj University, Madurai, Tamil Nadu, India

OPEN ACCESS

Received: 23/11/2023
Accepted: 14/04/2025
Published: 30/05/2025

Citation: Alagukumar S, Kathirvalavakumar T (2025) Impacts of Mutual Information and Discretization Methods in Associative Classifications. Indian Journal of Science and Technology 18(20): 1565-1577. https://doi.org/10.17485/IJST/v18i20.2983

* **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

Abstract

Objective: Data Pre-processing plays a vital role in associative classification to influence the learning outcome. Feature selection and discretization methods contribute more in implementing associative classification especially on gene expression data. The discretization method determines discrete intervals and these intervals are impacted on number of itemsets to create a transaction format for generating class association rules. This paper aims to find a suitable combination of feature selectionand discretization methods to generate association rules for classifying the cancer gene expression data. **Methods:** Three mutual information based features selection methods joint impurity filter (JIM), joint mutual information filter (JMI) and third-order joint mutual information filter(JMI3) are considered and five supervised discretization methods minimum description length principles (MDLP), class-attribute inter dependence maximization (CAIM), class-attribute contingency coefficient (CACC), an autonomous discretization algorithm (AMEVA) and chi-squared Merge (Chi-Merge) are considered for selecting the best combination. This combination is used in CMAR and CBA associative classification. **Findings:** CMAR gives 100% accuracy for the cancer gene expression datasets GSE15781 and GSE2685 for any combination of the assumed filter method and supervised discretization method. The considered evaluation metrics brier score, F-beta score, F1 score and AUC curve are also shown that CMAR provides better results when the information based filter method is combined with the supervised discretization. **Novelty**: Proposed method is used to find the informative genes using the opt mutual information method and discretization method to predict the gene is affected by cancer or not. The proposed method diagnoses diseases from microarray gene expression data using associative classification techniques for classifying the cancer gene expression data.

Keywords: Gene expression; Mutual information; Discretization; Associative classification; CBA; CMAR

1 Introduction

Machine learning techniques have been used for knowledge discovery in biological systems and popularly association rules mining is used. In recent decades, association rule mining is used for classification called associative classification. Associative classification techniques generate the class association rules in the form of (antecedent)→(consequent) statements. The antecedent represents the attribute values, and the consequent represents the class label[(1)]. Before generating the class association rules, significant features are to be selected and converted into discrete intervals. Initially raw dataare filtered using mutual information and are converted into discrete intervals using discretization. Then the discretized data sets are transformed into transactional data format. Number oftransactions in the transaction dataset is determined by the discretization methods. Number of class association rule is determined based on number of transactions. Transaction datasets are used to generate class association rules. If the number of transaction is increased then generating class association rules and execution time of the classification areincreased. If the number of transactions is decreased then association rules and execution time of the classification modelare decreased. Selecting opt mutual information and discretization methods for the labeled dataset isessential to create the class transaction data without loss of information. This paper mainly focuses on two pre-processing methods namely mutual information based filter methods and discretization methods for the associative classification to identify the informative genes and to predict the gene is affected by cancer or not. Data preprocessing is an important step in associative classification to optimize the training time and outcome[(2)]. Results of the classification model are impacted by the quality of the input features[(3)]. Gakii and Rimiru[(4)] have analyzed cancer related genes using feature selection and association rule mining. Identifying informative genes from large amounts of data is always a challenging one. Feature selection reduces the complex data into a data with lesser number of features. Robindro et al.[(5)] have introduced joint mutual information with class relevance for selecting mutual information.The joint mutual information with class relevance method was compared with other filter-based methods on various classifications algorithms and found its better performance. Salem et al.[(6)] have proposed a fuzzy joint mutual information based on ideal vector to extract significant features with minimum computational time and improved the classification performance. Bommert et et al.[(7)] have reviewed 16 high dimensional classification data sets and 22 filter methods on machine learning algorithms and concluded that the filter based methods can be combined with any machine learning model and it heavily reduce the run time of machine learning algorithms. Discretization is a data preprocessing that also plays a vital role in the associative classification[(8)]. Discretization techniques are classified as supervised and unsupervised methods. The unsupervised discretization converts the continuous data into discrete intervals using user specification but the supervised discretization uses class label to convert the continuous data into discrete intervals. Varghese and Sundari[(9)] have compared the performance of random forest classifiers after discretization preprocessing. They have pointed out that discretization improves the accuracy of random forest classification algorithms. Putri et al.[(10)] have analyzed the performance of equal width and equal frequency unsupervised discretization methods on data features in classification process and significantly improved the accuracy in classification models. It has been observed from the existing studies that selecting subset of mutual information and preparing best discretization intervals are essential for the associative classification. Based on the discretization, number of class intervals varied. This work aims to generate minimal number of discretization intervals for gene expression for better classification. This paper suggests a method to determine filter based feature selection method with the supervised discretization for the effective associative classification for the gene expression data. Methods are explained in section 2 and section 3 describes experimental results and discussion.

2 Methodology

Performance of the associative classifications CBA (Classification Based on Associations) and CMAR (Classification based on Multiple Association Rules) with the three different mutual information based features selection methods and five different supervised discretization methods are analyzed to find the combination which is better in identifying the informative genes and predicting whether the gene is affected by cancer or not. Figure 1 depicts the schematic diagram of the proposed associative classification. The proposed method has three phases, mutual information filter, discretization and associative classification.

2.1 Mutual Information Filter

Entropy and mutual information in the information theory are widely used for feature selection[(11)],[(12)],[(13)]. In this work, informative gene expression is selected by using any one of the mutual information based feature selection methods joint mutual information filter, joint impurity filter and third-order joint mutual information.

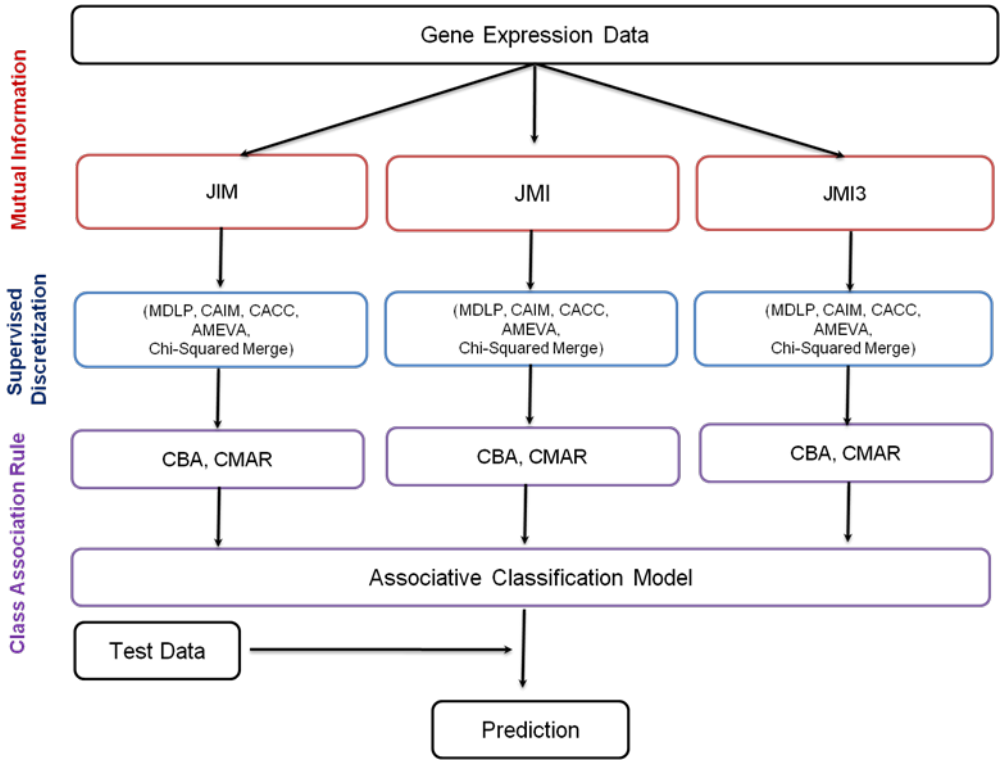

Fig 1. Schematic representation of proposed associative classification

2.1.1. Joint Mutual Information Filter

The joint mutual information filter[11] uses maximal mutual information. It starts by finding maximal mutual information for each feature of each class label. Then sorts the features in descending order based on the maximal mutual information which is found. Required number of top features can be selected for further processing. The JMI mutual informative features are selected using **Eq. (1)**, where C, S, W and D represent the available features, set of features, selected features and class labels respectively.

$$JMI(C) = \sum_{W \in S}(C, W; D) \tag{1}$$

2.1.2. Joint Impurity Filter

Joint impurity filter[11] is similar to JMI filter method but in addition it uses Gini impurity score to select significant features. It executes faster. After finding maximal mutual information for each feature, impurity score for each feature is calculated based on the maximal mutual information. Now features are sorted in descending order based on impurity score. Required number of top features can be selected for generating association rules. The impurity score is calculated using **Eq. (2)**. $G(C, D)$is the Gini impurity gain for$D$where P_{cd} is a distribution of the joint states of C and D , whileP_c and P_d represent respective marginal values.

$$G(C, D) = \sum_{cd} \frac{P_{cd}^2}{P_c} - \sum_{cd} P_d^2 \tag{2}$$

2.1.3. Third-order Joint Mutual Information Filter

JMI3 method calculates the mutual information between every consecutive two features based on joint probability distribution. It works like joint mutual information filter except that the mutual information are calculated for every consecutive two features and the class label D. Then, it greedily adds features with a maximal value to filter the gene expression using **Eq. (3)**, where S represents set of available features, Crepresents the two selected gene features, D represente the class label and U, W are the set

of selected features.

$$J(C) = \frac{1}{2} \sum_{(U,W) \in S} I(C,U,W;D) \tag{3}$$

2.2 Supervised Discretization

In this phase, gene expression values are replaced with distinct intervals using a discretization technique. Five supervised discretization methods minimum description length principle, class-attribute inter dependence maximization, class-attribute contingency coefficient, autonomous discretization algorithm and Chi-merge discretization are considered for choosing the best.

2.2.1. Minimum Description Length Principle

Minimum description length principle (MDLP) follows top down approach[14],[15],[16]. This method uses class information to split the data into intervals and choose useful cut-points. The MDLP process is explained in the Algorithm 1.

Algorithm 1: MDLP Discretization

Input: Mutual informative gene expression

Output: Discretized Gene Expression

Process:

Begin

Step1: Read the mutual informative gene expression

Step2: Sort the gene expression values

Step3: Calculate entropy H(X)for the gene expression data using **eq. (4)**, where P_c represents the probability of class c. $\log_2$ represents the information encoded in bits.

$$Info(S) = -\sum P_c \log_2 (P_c) \tag{4}$$

Step4: Search a suitable cut point with the lowest entropy

Step5: Spilt the continuous gene expression values according to the cut point using minimum Descriptionlength principle

Step6: Repeat steps 4-5 until all the continuous values are discretized.

End

2.2.2. Class-Attribute Interdependence Maximization

The CAIM discretization algorithm is follows a top down approach[16]. The CAIM algorithm maximizes the class attribute interdependence and used to find the minimal number of discrete intervals. This discretization process is explained in the Algorithm 2.

Algorithm 2: CAIM Discretization

Input: Mutual informative gene expression

Output: Discretized Gene Expression

Process:

Begin

Step1: Read the mutual informative gene expression

Step2: Sort the gene expression values

Step3: CAIM is calculated between the class and its attributes using the **Eq. (5)** to spilt the continuous gene expression into discrete intervals

$$CAIM = \frac{\sum_{c=1}^{n} \frac{m_c^2}{D+c}}{n} \tag{5}$$

Where $c = 1, 2, ..., n, m_c^2$ is the maximum value within the c^{th} column of the matrix. $D + c$ is the total number of continuous values of gene attribute that are within the interval.

Step4: Repeat step 3 until all the continuous values are discretized

End

2.2.3. Class-Attribute Contingency Coefficient

The CACC discretization algorithm[16] uses a class label for discretizing the data. It works on top down approach and uses contingency coefficient to find the number of discrete intervals. The CACC discretization process is explained in the algorithm 3.

Algorithm 3: CACC Discretization

Input: Mutual informative gene expression

Output: Discretized Gene Expression

Process:

Begin

Step1: Read the mutual informative gene expression

Step2: Sort the gene expression values

Step3:Contingency coefficient value is calculated between class variable and attributes using the **Eq. (6)** to spilt the continuous gene expression into discrete intervals

$$CACC = \sqrt{\frac{y}{y+S}} \tag{6}$$

where $y = \frac{\chi^2}{log(n)}$, $S$is the total number of attributes, n is the number of discretized intervals and y is the chi-square statistic.

Step4: Repeat step 3 until discretize all the continuous values

End

2.2.4. Autonomous Discretization Algorithm

An autonomous discretization [AMEVA] is based on the Chi-squared test. AMEVA[16] follows top down approach. The AMEVA uses chi-squared statistic for optimal discretization and provides less number of intervals.The AMEVAdiscretization process is explained in the algorithm 4.

Algorithm 4: AMEVA Discretization

Input: Mutual informative gene expression

Output: Discretized Gene Expression

Process:

Begin

Step1: Read the mutual informative gene expression

Step2: Sort the gene expression values

Step3: AMEVA is calculated using the **Eq. (7)** to spilt the continuous gene expression into discrete intervals,where d is a number of intervals, c is a number of classes.

$$Ameva(d) = \frac{\chi^2(d)}{d*(c-1)} \tag{7}$$

Step 4: Repeat step 3 until discretize all the continuous values

End

2.2.5. Chi-Merge Discretization

Chi-Merge discretization[16] follows abottom-up approach. It uses the chi-squared statistic to determine whether the class frequencies of adjacent intervals are distinctly different or similar, to merge them into a single interval. The Chi-Merge discretization process is explained in the algorithm 5.

Algorithm 5: Chi-Merge Discretization

Input: Mutual informative gene expression

Output: Discretized Gene Expression

Process:

Begin

Step1: Read the mutual informative gene expression

Step2: Sort the gene expression values

Step3: Compute the χ^2 value for each pair of adjacent intervals using **Eq. (8)**.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{8}$$

where k represents number of classes, A_{ij} represents number samples in the i^{th} interval of the j^{th} class. E_{ij} represents the expected $\frac{(R_i * C_j)}{N}$ frequency of A_{ij},R_i represents number of samples in the i^{th} interval and C_j represents number of samples in the i^{th} class and Nrepresents total number of samples on the two intervals.

Step4: Merge the pair of adjacent intervals with the lowest χ^2 value.

Step5: Repeat steps 3 and 4 until all the continuous values are discretized.

2.3 Associative Classification

In this phase, two associative classification methods namely CBA and CMAR are used to classify the data. Discretized features are fed into the CBA[17] and CMAR[18] classifiers to generate class association rules. CBA algorithm follows the apriori based approach and CMAR follows the FP-growth based approach. Associative classification methods generate the association rules in the form of (itemsets)→(class label) statements. A class association rule is an association rule[19] where the antecedent of the rules contains gene items and consequent of the rule is class label for classification. In the associative classification, first discretized data are transformed into the transaction data. Then the frequent itemsets are generated from the transaction data. Finally, class association rules are generated from the frequent itemsets.

2.3.1. Accuracy

Associative Classification prediction accuracy A_C is calculated using **Eq. (9)** where T_p is the True Positive which represents correctly detected diseased genes of cancer patients. F_n is the False Negative which represents wrongly detected diseased genes of cancer.

$$A_C = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{9}$$

3 Results and Discussion

Experiments are carried out with R Tool. The microarray data colorectal cancer with the accession number GSE15781 and gastric cancer dataset with the accession number GSE2685 are downloaded from the Gene Expression Omnibus (GEO) database[20]. GSE15781 dataset consists of the samples with 13 cancer tissues and 10 normal tissues of colorectal cancer patients. GSE2685 consists of samples with 22 primary cancer tissues and 8 noncancerous tissues of gastric cancer patients. Table 1 describes the data set.

Table 1. Overview of the microarray datasets

Dataset	GSE ID	No. of gene Features	No. of samples	No. of cancer samples	No. of normal samples
Colorectal Cancer	GSE15781	32879	23	13	10
Gastric Cancer	GSE2685	7129	30	22	8

Raw data consist of thousands of gene features are filtered using mutual information methods. Top 50 genes associated with colorectal cancer and gastric cancer are identified using the JIM, JMI and JMI3 mutual information methods. The selected mutual informative data are fed into the five different discretization methods. Table 2 represents number of transaction sets created from three mutual information methods and five discretization methods. It is observed from Table 2 that the MDLP discretization method provides less number of itemsets for the JIM, JMI and JMI3 filter methods.

3.1 Performance Analysis

Performance of the proposed work is evaluated using Accuracy, Brier Score, F-Beta score and F1 Score[21]. The Brier score measures the mean squared difference between the predicted probability and the actual outcome. The Brier score is used

Table 2. Transaction and itemsets created from mutual Information and discretization methods

Dataset	Mutual Information Methods	K- Mutual Important Features	Discretization	No. of Transactions	No. of Itemsets
GSE15781	JIM	50	MDLP	18	89
	JIM	50	CAIM	18	202
	JIM	50	CACC	18	253
	JIM	50	AMEVA	18	221
	JIM	50	Chi Squared Merge	18	179
GSE15781	JMI	50	MDLP	18	89
	JMI	50	CAIM	18	203
	JMI	50	CACC	18	256
	JMI	50	AMEVA	18	225
	JMI	50	Chi Squared Merge	18	185
GSE15781	JMI3	50	MDLP	18	84
	JMI3	50	CAIM	18	204
	JMI3	50	CACC	18	256
	JMI3	50	AMEVA	18	224
	JMI3	50	Chi Squared Merge	18	167
GSE2685	JIM	50	Entropy	24	100
	JIM	50	CAIM	24	202
	JIM	50	CACC	24	211
	JIM	50	AMEVA	24	206
	JIM	50	Chi Squared Merge	24	122
GSE2685	JMI	50	Entropy	24	98
	JMI	50	CAIM	24	202
	JMI	50	CACC	24	224
	JMI	50	AMEVA	24	204
	JMI	50	Chi Squared Merge	24	125
GSE2685	JMI3	50	Entropy	24	96
	JMI3	50	CAIM	24	198
	JMI3	50	CACC	24	231
	JMI3	50	AMEVA	24	202
	JMI3	50	Chi Squared Merge	24	134

to identify the error rate of the associative classification model. The value of the Brier score is always between 0.0 and 1.0, where 0 represents a perfect model and 1 represents worst model. The F-beta score is the weighted mean of precision and recall. The score ranges between 0 and 1. The closer the 1 represents the best associative classification model. The F1 score is calculated as the mean of the precision and recall scores. F1-score ranges between 0 and 1. The closer to 1 represents the better associative classification model.

Table 3 represents the performance of the CBA and CMAR algorithm on three mutual information methods and five discretization methods. Performance is evaluated in terms of accuracy and time taken for classification when the support count is 20% and confidence is 100%. With the combination of MDLP discretization and JIM, JMI and JMI3 filter methods, CBA

Table 3. Accuracy and Execution Time with support count 20% and confidence 100%

Dataset	Mutual Information Methods	Discretization	No. of Transactions	No. of Itemsets	CBA		CMAR	
					Accuracy	Time (Sec.)	Accuracy	Time (Sec.)
GSE15781	JIM	MDLP	18	89	100	3.53	100	0.34
		CAIM	18	202	100	2.68	100	0.33
		CACC	18	235	60	1.84	100	0.34
		AMEVA	18	221	80	1.99	100	0.34
		Chi Squared Merge	18	179	80	1.49	100	0.34
GSE15781	JMI	MDLP	18	89	100	3.19	100	0.34
		CAIM	18	203	100	2.71	100	0.33
		CACC	18	256	100	1.83	100	0.34
		AMEVA	18	225	80	2.52	100	0.34
		Chi Squared Merge	18	185	20	1.66	100	0.34
GSE15781	JMI3	MDLP	18	84	100	3.2	100	0.81
		CAIM	18	204	60	2.68	100	0.33
		CACC	18	256	80	1.74	100	0.34
		AMEVA	18	224	40	2.14	100	0.35
		Chi Squared Merge	18	167	100	1.67	100	0.34
GSE2685	JIM	MDLP	24	100	66.67	5.02	100	9.54
		CAIM	24	202	83.33	2.22	100	1.12
		CACC	24	211	83.33	2.57	100	1.12
		AMEVA	24	206	83.33	2.36	100	1.09
		Chi Squared Merge	24	122	66.67	4.13	100	15.99
GSE2685	JMI	MDLP	24	98	16.67	4.94	100	9.51
		CAIM	24	202	16.67	2.89	100	1.1
		CACC	24	224	16.67	2.77	100	1.09
		AMEVA	24	204	16.67	3.03	100	1.1
		Chi Squared Merge	24	125	83.33	4.12	100	15.75
GSE2685	JMI3	MDLP	24	96	83.33	4.99	100	7.66
		CAIM	24	198	83.33	2.43	100	0.95
		CACC	24	231	100	2.49	100	1.11
		AMEVA	24	202	83.33	2.68	100	0.96
		Chi Squared Merge	24	134	83.33	4.14	100	16.92

algorithm provides 100% accuracy for the GSE15781 dataset but CMAR algorithm provides 100 % accuracy for the GSE15781 and GSE2685 gene expression datasets.

Table 4. Accuracy and Execution Time with support count 50% and confidence 100%

Dataset	Mutual Information Methods	Discretization	No. of Transactions	No. of Itemsets	CBA		CMAR	
					Accuracy	Time (Sec.)	Accuracy	Time (Sec.)
GSE15781	JIM	MDLP	18	89	100	0.73	100	0.31
		CAIM	18	202	100	0.38	100	0.31
		CACC	18	253	100	0.36	100	0.33
		AMEVA	18	221	100	0.35	100	0.33
		Chi Squared Merge	18	179	100	0.35	100	0.33
GSE15781	JMI	MDLP	18	89	100	0.75	100	0.31
		CAIM	18	203	100	0.35	100	0.31
		CACC	18	256	100	0.35	100	0.33
		AMEVA	18	225	100	0.33	100	0.33
		Chi Squared Merge	18	185	100	0.34	100	0.33
GSE15781	JMI3	MDLP	18	84	100	0.84	100	0.8
		CAIM	18	204	100	0.34	100	0.36
		CACC	18	256	100	0.33	100	0.34
		AMEVA	18	224	100	0.33	100	0.35
		Chi Squared Merge	18	167	100	0.36	100	0.34
GSE2685	JIM	MDLP	50	100	50	2.15	100	9.34
		CAIM	50	202	50	2.26	100	1.08
		CACC	50	211	50	2.13	100	1.15
		AMEVA	50	206	50	2.11	100	1.13
		Chi Squared Merge	50	122	50	1.6	100	16.86
GSE2685	JMI	MDLP	50	98	100	1.99	100	9.31
		CAIM	50	202	83.33	2.2	100	1.07
		CACC	50	224	100	1.63	100	0.94
		AMEVA	50	204	83.33	1.76	100	1.09
GSE2685	JMI3	Chi Squared Merge	50	125	100	1.55	100	14.78
		MDLP	50	96	100	1.51	100	8.02
		CAIM	50	198	83.33	2.08	100	1.07
		CACC	50	231	83.33	1.07	100	1.1
		AMEVA	50	202	83.33	1.23	100	0.99
		Chi Squared Merge	50	134	100	0.73	100	16.14

Table 4 represents the performance of the CBA and CMAR algorithm for the support count 50% and the confidence 100%. With the combination of MDLP discretization and JIM, JMI and JMI3 filter methods, CBA provides 100% accuracy for the GSE15781 dataset but CMAR algorithm provides 100% accuracy for the GSE15781 and GSE2685 datasets.

Table 5 represents the performance of the combination of the discretization method MDLP with the filter methods JIM, JMI and JMI3. It is observed that CBA provides best Brier Score, F-Beta Score and F1 score for the GSE15781 dataset. But the CMAR algorithm provides best Brier Score, F-Beta Score and F1 score for the datasets GSE15781 and GSE2685.

Table 5. Evaluation Metrics of the Proposed Method with support count 20% and confidence 100%

Dataset	Mutual Information + Discretization	CBA				CMAR			
		Accuracy	Brier Score	F1 Score	F-Beta Score	Accuracy	Brier Score	F1 Score	F-Beta Score
GSE15781	JIM + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + CAIM	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + CACC	60	0.40	0.66	0.75	100	0.0	1.0	1.0
	JIM + AMEVA	80	0.20	0.85	0.83	100	0.0	1.0	1.0
	JIM + Chi Squared Merge	80	0.20	0.85	0.83	100	0.0	1.0	1.0
GSE15781	JMI + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + CAIM	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + AMEVA	80	0.20	0.85	0.83	100	0.0	1.0	1.0
	JMI + Chi Squared Merge	20	0.80	0.33	0.41	100	0.0	1.0	1.0
GSE15781	JMI3 + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + CAIM	60	0.40	0.66	0.75	100	0.0	1.0	1.0
	JMI3 + CACC	80	0.20	0.85	0.83	100	0.0	1.0	1.0
	JMI3 + AMEVA	40	0.60	0.40	0.62	100	0.0	1.0	1.0
	JMI3 + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0
GSE2685	JIM + MDLP	66.67	0.33	0.75	0.88	100	0.0	1.0	1.0
	JIM + CAIM	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JIM + CACC	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JIM + AMEVA	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JIM + Chi Squared Merge	66.67	0.33	0.75	0.88	100	0.0	1.0	1.0
GSE2685	JMI + MDLP	16.67	0.83	0.29	0.38	100	0.0	1.0	1.0
	JMI + CAIM	16.67	0.83	0.29	0.38	100	0.0	1.0	1.0
	JMI + CACC	16.67	0.83	0.29	0.38	100	0.0	1.0	1.0
	JMI + AMEVA	16.67	0.83	0.29	0.38	100	0.0	1.0	1.0
	JMI + Chi Squared Merge	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
GSE2685	JMI3 + MDLP	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + CAIM	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + AMEVA	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + Chi Squared Merge	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0

Table 6 represents the combination of the five discretization methods with the filter methods JIM, JMI and JMI3. It has been observed from the table that CBA provides best Brier Score, F-Beta Score and F1 score for the GSE15781 dataset. But the CMAR algorithm provides best Brier Score, F-Beta Score and F1 score for the datasets GSE15781 and GSE2685.

Figure 2 depicts the AUC curve of the proposed methods on the colorectal and gastric cancer data sets. It is observed from the experiments that MDLP supervised discretization provides less number of intervals and is highly impacted on class association rules to give good results on accuracy, Brier score, F-beta score, F1 score and AUC curve.

The Table 7 compares the proposed work with the existing methods [10] for the cancer gene expression datasets GSE15781 and GSE2685. The existing methods use unsupervised discretization methods namely equal width interval binning, equal frequency, and K-Means clustering discretizations to discrete the intervals based on a user specified k value. Number of intervals increased

Table 6. Evaluation Metrics of the Proposed Method with support count 50% and confidence 100%

Dataset	Mutual Information + Discretization	CBA				CMAR			
		Accuracy	Brier Score	F1 Score	F-Beta Score	Accuracy	Brier Score	F1 Score	F-Beta Score
GSE15781	JIM + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + CAIM	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + AMEVA	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JIM + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0
GSE15781	JMI + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + CAIM	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + AMEVA	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0
GSE15781	JMI3 + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + CAIM	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + AMEVA	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0
GSE2685	JIM + MDLP	50	0.5	0.57	0.76	100	0.0	1.0	1.0
	JIM + CAIM	50	0.5	0.57	0.76	100	0.0	1.0	1.0
	JIM + CACC	50	0.5	0.57	0.76	100	0.0	1.0	1.0
	JIM + AMEVA	50	0.5	0.57	0.76	100	0.0	1.0	1.0
	JIM + Chi Squared Merge	50	0.5	0.57	0.76	100	0.0	1.0	1.0
GSE2685	JMI + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + CAIM	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI + CACC	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI + AMEVA	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0
GSE2685	JMI3 + MDLP	100	0.0	1.0	1.0	100	0.0	1.0	1.0
	JMI3 + CAIM	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + CACC	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + AMEVA	83.33	0.16	0.88	0.95	100	0.0	1.0	1.0
	JMI3 + Chi Squared Merge	100	0.0	1.0	1.0	100	0.0	1.0	1.0

or decresed based on the user specified k values and information loss also occurs. The proposed method uses five different supervised discretization methods to discretize the gene features into set of gene intervals using the class information. It creates transaction data without any loss of information to generate class association rules.

It is observed from Table 7 that when the user specifies k-value as 2 in the unsupervised discretization methods namely equal width interval, equal frequency and K-Means cluster then Naïve Bayes classifer gives 91.67% accuracy, but it gives 100% classification accuracy when the k-value is selected as 3 for the dataset GSE15781. But for the dataset GSE2685, 80% accuracy is obtained when the interval is 2 but 100% accuracy is obtained when the interval is 3. But any combination of the supervised discretization and CMAR provides 100% of accuracy without specifying the k-intervals for the gene expression data sets GSE15781 and GSE2685. Similarly the combination of MDLP discretization and CBA provides 100% of accuracy without

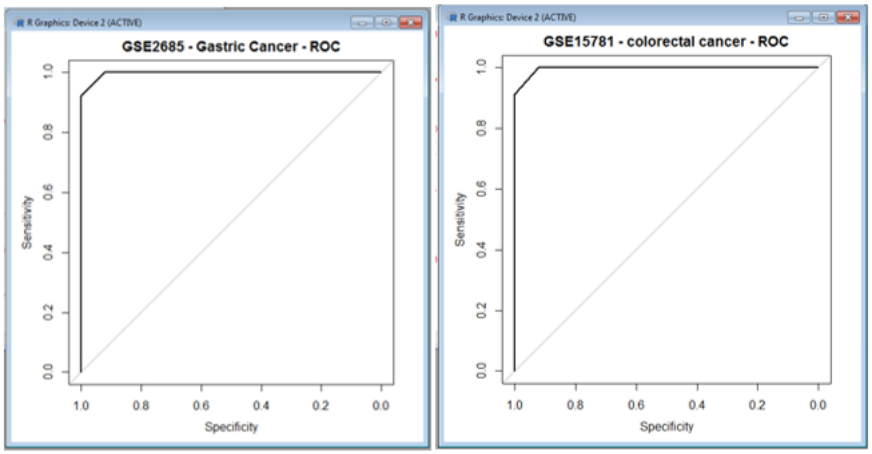

Fig 2. The Area Under the Curve for GSE2685 and GSE15781

Table 7. Comparision of proposed work with the existing methods

Dataset	Discretization + Classification	Accuracy in %	
		User Specified K-Intervals (K=2)	User Specified K-Intervals (K=3)
GSE15781	Equal Width Interval + Naïve Bayes [10]	91.67	100
	Equal Frequecy + Naïve Bayes [10]	91.67	100
	K-Means Cluster + Naïve Bayes	91.67	100
	Proposed Method (MDLP discretization+ CBA)	**Provides 100% accuracy without specify the K user intervals**	
	Proposed Method (any combination of the supervised discretization method + CMAR)	**Provides 100% accuracy without specify the K user intervals**	
GSE2685	Equal Width Interval + Naïve Bayes [10]	80	100
	Equal Frequecy + Naïve Bayes [10]	80	100
	K-Means Cluster + Naïve Bayes	80	100
	Proposed Method (any combination of the supervised discretization method + CMAR)	**Provides 100% accuracy without specify the K user intervals**	

specifying the k-intervals for the gene expression data GSE15781. It is observed from the Table 7 that the proposed model performs well than the existing model [10]. It is observed that supervised discretization finds optimal gene intervals which leads to build more accurate models for gene expression data and improves the classification performance.

4 Conclusion

Discretization and feature selection methods are highly impacted on the outcome of the associative classification. Three mutual information based features selection methods namely joint impurity filter (JIM), joint mutual information filter (JMI) and third-order joint mutual information filter(JMI3) and five supervised discretization methods namely minimum description length principles (MDLP), class-attribute interdependence maximization (CAIM), class-attribute contingency coefficient (CACC), an autonomous discretization algorithm (AMEVA) and chi-squared Merge (ChiMerge) are tested with different combinations. When a filter method JIM or JMI or JMI3 is associated with MDLP discretization, CBA provides 100%

classification accuracy for GSE15781. For any combination of the assumed filter method and supervised discretization method, CMAR associative classification gives 100% accuracy for the cancer gene expression datasets GSE15781 and GSE2685. The considered evaluation metrics Brier score, F-beta score, F1 score and AUC curve are also shown that CMAR provides good results when the information based filter method is combined with the supervised discretization. Proposed method is used to find the informative genes to predict the gene is affected by cancer or not. The proposed method diagnoses diseases from microarray gene expression data using associative classification techniques. The proposed method uses mutual information filter methods, discretization techniques and generates class association rules for classifying the cancer gene expression data. The proposed method produces feasible number of distinct gene intervals which improves the classification accuracy and minimize computation time.

References

1) Seyfi M, Xu Y, Nayak R. DAC: Discriminative Associative Classification. *SN Computer Science*. 2023;4(4):401. https://doi.org/10.1007/s42979-023-01819-9.
2) Tran THA, Lara M. Influence of discretization granularity on learning classification models. *BNAIC/BeNeLearn Joint International Scientific Conferences on AI and Machine Learning*. 2022. Available from: https://bnaic2022.uantwerpen.be/BNAICBeNeLearn_2022_submission_8652.
3) Gonzalez-Lopez J, Ventura S, Cano A. Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Systems*. 2020;188:105052. https://doi.org/10.1016/j.knosys.2019.105052.
4) Gakii C, Rimiru R. Identification of cancer related genes using feature selection and association rule mining. *Informatics in Medicine Unlocked*. 2021;24:100595. https://doi.org/10.1016/j.imu.2021.100595.
5) Robindro K, Clinton UB, Hoque N, Bhattacharyya DK. JoMIC: A joint MI-based filter feature selection method. *Journal of Computational Mathematics and Data Science*. 2023;6:100075. https://doi.org/10.1016/j.jcmds.2023.100075.
6) Salem OA, Liu F, Chen YPP, Hamed A, Chen X. Fuzzy joint mutual information feature selection based on ideal vector. *Expert Systems with Applications*. 2022;193:116453. https://doi.org/10.1016/j.eswa.2021.116453.
7) 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;143:106839. https://doi.org/10.1093/bib/bbab354.
8) Toulabinejad E, Mirsafaei M, Basiri A. Supervised discretization of continuous-valued attributes for classification using RACER algorithm. *Expert Systems with Applications*. 2023;p. 121203. https://doi.org/10.1016/j.eswa.2023.121203.
9) Varghese IS, Sundari RG. The performance enhancement through discretization of variables in classification using random forest as classifier. *AIP Conference Proceedings*. 2023. https://doi.org/10.1063/5.0139159.
10) Putri PAR, Prasetiyowati SS, Sibaroni Y. The Performance of the Equal-Width and Equal-Frequency Discretization Methods on Data Features in Classification Process. *Sinkron: jurnal dan penelitian teknik informatika*. 2023;8(4):2082–2098. https://doi.org/10.33395/sinkron.v8i4.12730.
11) Kursa MB. Praznik: High performance information-based feature selection. *SoftwareX*. 2021;16:100819. https://doi.org/10.1016/j.softx.2021.100819.
12) Ceglia N, Sethna Z, Freeman SS, Uhlitz F, Bojilova V, Rusk N, et al. Identification of transcriptional programs using dense vector representations defined by mutual information with GeneVector. *Nature Communications*. 2023;14:4400. https://doi.org/10.1038/s41467-023-39985-2.
13) Anh CT, Kwon YK. Mutual Information Based on Multiple Level Discretization Network Inference from Time Series Gene Expression Profiles. *Applied Sciences*. 2023;13(21):11902. https://doi.org/10.3390/app132111902.
14) Senozan H, Soylu B. A flexible non-monotonic discretization method for pre-processing in supervised learning. *Pattern Recognition Letters*. 2024;181:77–85. https://doi.org/10.1016/j.patrec.2024.03.024.
15) Mudumba B, Kabir MF. Mine-first association rule mining: An integration of independent frequent patterns in distributed environments. *Decision Analytics Journal*. 2024;10:100434. https://doi.org/10.1016/j.dajour.2024.100434.
16) Yılmaz KAYA, Tekin R. Comparison of discretization methods for classifier decision trees and decision rules on medical data sets. *Avrupa Bilim ve Teknoloji Dergisi*. 2022;(35):275–281. Available from: https://dergipark.org.tr/tr/download/article-file/2278969.
17) Ma BLWHY, Liu B, Hsu Y. Integrating classification and association rule mining. *Proceedings of the fourth international conference on knowledge discovery and data mining*. 1998;50:55. Available from: https://dl.acm.org/doi/10.5555/3000292.3000305.
18) Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules. *Proceedings 2001 IEEE international conference on data mining*. 2001;p. 369–376. Available from: https://www.cs.sfu.ca/~jpei/publications/cmar.pdf.
19) Wu W, Wang S, Liu B, Shao Y, Xie W. A novel software defect prediction approach via weighted classification based on association rule mining. *Engineering Applications of Artificial Intelligence*. 2024;129:107622. https://doi.org/10.1016/j.engappai.2023.107622.
20) Xu H, Ma Y, Zhang J, Gu J, Jing X, Lu S, et al. Identification and Verification of Core Genes in Colorectal Cancer. *Biomed Res Int*. 2020. https://doi.org/10.1155/2020/8082697.
21) Deng Q, Li S, Zhang Y, Jia Y, Yang Y. Development and validation of interpretable machine learning models to predict distant metastasis and prognosis of muscle-invasive bladder cancer patients. *Scientific Reports*. 2025;15(1):11795. https://doi.org/10.1038/s41598-025-96089-1.