A New Gamma Mining Procedure Case Study: Breast Cancer Tumor Diagnosis

S. Abd El-Badie

IT Sector, Man Power & Immigration Directorate, Tanta, Egypt

Abstract

In this paper, an innovative Statistical Data Mining (SDM) technique is proposed using Gamma Mining Procedure (GMP) contributing a new classifier & predictor by applying very effective stages on the training and testing data depending on Gamma (G) correlation matrix and Gamma absorption process. Linking the previous stages with the Misclassification Error (M_{Error}) as a precision measure for obtaining a new classifier and a new predictor, then using the novel predictor for attributes and objects mining of the test data. Applying the last GMP stage by using the contributed predictor attributes with Naive Bayes technique for prediction. The proposed GMP technique is applied and examined on a Breast Cancer Tumor diagnosis to demonstrate its applicability. Two SDM validation tools are used with the new SDM technique, the 1st versus cross validation with bootstrapping using Rapid Miner as a DM tool, and the 2nd versus two step cluster analysis, using SPSS Modeler.

Keywords

Gamma, Predictor, Classifier, Naïve Bayes, Rapid Miner.

1. Introduction

The wide variety of data gathering methods in the 21st century caused many multifaceted problems which are considered as a great challenge to achieve the greatest benefit from this data. Statistical Data Mining (SDM) techniques [2, 8,9,15,16] have long been employed to find the decision core in any problem type. There are many techniques for carrying out classification and prediction. The data type is a strong effective factor for deciding which SDM will be used for the analysis such as (Covid-19 Data, Climate Change Data, Space Change Data…etc.). The data type of our application is ordinal categorical data; GMP depends on using Gamma correlation coefficient [5,17] which is the most appropriate coefficient with this type of data. In this paper a novel Statistical Data Mining (SDM) technique depending on new initiated classifier and contributed predictor using Gamma Mining Procedure (GMP). The proposed procedure is applied on a real-life data of one of the most vital problems [1,2], which is Breast Cancer Tumor Diagnosis.

The interesting sequence of this paper is organized as follows; the next section represents the materials & methods [1,2,3,4] which are, Breast Cancer Data characteristics, Gamma (G) correlation coefficient, Misclassification Error (M_{Error}) and Naïve Bayes procedure. After that the practical work of the paper is divided into 7 stages, the 1st four stages are the manual work of this paper. While, the 5th stage is the flow chart of GMP systematic structure stages using Gamma correlation matrix, Gamma classes, Gamma classifiers, Gamma averages classes, Gamma averages classifiers and Gamma average absorption mining. Linking all the previous stages with the Misclassification Error (M_{Error}) and M_{Error} average as precision measures for initiating *a new classifier and a new predictor*. From the other hand, the software validation of GMP procedure is represented in the 6th stage using RapidMiner [11,13]. And the 7th stage using SPSS Modeler.

DOI: 10.5121/ams.2022.9402

Finally, the conclusion and the future work. The tables and the figures of this paper are represented in separated sections at the end of the paper.

2. GMP MATERIALS & METHODS

Splitting data into training and testing sets is an important part of evaluating data mining models [2,8,14], most of the data is used for training, and a smaller portion of the data is used for testing. After a model has been processed by using the training set, test the model by making predictions against the test set [6,12]. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct. In this section, the description and the coding of our medical data real life application material will be introduced. From the other hand the methods which are used in this paper definitions and relations will be described in this section.

2.1. Breast Cancer Data: Characteristic

The material of the study [1,2,4] described by a breast cancer data which included 30 cases of female breast tumors received from the Pathology department at Tanta Cancer center. The age of the patients ranged from 20-55 years, 20 of them are training data and 10 will be the test data. All the patients were subjected to clinical studies by studying the files of the patients regarding age and reviewing the slides of the diagnosis of each tumor. Table (1) present the breast cancer data description, Table (2) and Table (3) introduce the training and the test coded data. Where, $\{A_1, A_2, A_3 \ldots A_{10}\}$ are the conditional attributes, $\{P_1,P_2,P_3,\ldots,P_{20}\}$ in Table (2) are the Breast Cancer training data objects, $\{P_1,P_2,P_3,\ldots,P_{10}\}$ in Table (3) as the Breast Cancer test data objects and the diagnosis attribute (D) as the decision attribute.

2.2. Gamma Correlation Coefficient (G)

Gamma correlation coefficient [5,17] is a measure of rank correlation. In other words, the similarity of the orderings of the data when ranked by each of the quantities. It measures the strength of association of the cross tabulated data when both variables are measured at the ordinal level. It makes no adjustment for either table size or ties. The Gamma statistic computes the level of association between the two variables based on two sums that are obtained by constructing a cross tabulation and applying two formulas. The first sum represents the number of Agreements N_s. This sum is determined by identifying the number of cases that are ranked in the same relative position on both variables. The second sum represents the number of Inversions N_d. This sum is determined by identifying the number of cases that ranked differently on the two variables (higher on one and lower on another). The estimate of Gamma depends on two quantities:

- **N_S :** No. of Agreements (The number of cases that are ranked in the same relative position on both variables)

- **N_d:** No. of Inversions (The number of cases that ranked differently on the two variables.)

Where "ties" are dropped, that is cases where either of the two variables in the pair are equal. Then

$$\mathbf{G = \frac{N_s - N_d}{N_s + N_d}} \qquad \textbf{(1)}$$

This statistic can be regarded as the maximum likelihood estimator for the theoretical quantity γ, where

$$\gamma = \frac{\mathbf{P_s - P_d}}{\mathbf{P_s + P_d}} \qquad \mathbf{(2)}$$

and where P_s and P_d are the probabilities that a randomly selected pair of observations will place in the same or opposite order respectively, when ranked by both variables.

To compute G, begin with the construction of a cross tabulation that represents the observed values for each case under consideration on both of variables. When constructing the cross tabulation, the highest ranking should be at the top among the rows and at the left among the columns as in the following example and it is one case of our Breast Cancer data.

2.3. Misclassification Error (M_{error})

The Misclassification Error [2,8] is considered one of the most important measures for the classification precision. The following formula represents the rule of M_{Error}. Where,

- A_i: The requested attribute error.
- j: The requested case of the dependent attribute.

$$\mathbf{M_{Error}(A_i) = 1 - Max_j P(j/A_i)} \qquad \mathbf{(3)}$$

The value of the M_{Error} is maximum when records are equally distributed among all classes, implying least remarkable information and it takes the minimum value “0” when all records belong to one class, implying most remarkable information.

2.4. Naïve Bayes Procedure

Naïve Bayes technique [3, 7] is one of the most effective SDM classifiers and predictors depending on the concept of Bayes theorem. It is more suitable when the dimensionality of the input data is high regardless of its simplicity.

Naïve Bayes classifiers solve the problems with an arbitrary number of independent variables whether continuous or categorical. For a set of variables A = {A_1, A_2, A_3… A_n}, constructing the posterior probability for the event d_j among a set of possible outcomes D = {d_1, d_2, $d_{3...}d_n$} i.e. A are the predictors and D is the set of decision levels existing in the dependent variable. Using Bayes’ rule, The Predicted Posterior Probability is given by;

$$\mathbf{P(d_j/A_1, A_2, A_3, \ldots, A_n) = P(d_j) \prod_{i=1}^{n} P(A_i/d_j)} \quad \mathbf{(4)}$$

3. GMP 1STSTAGE: BEST GAMMA CLASSIFIER

The proposed SDM technique which named by; Gamma Mining Procedure (GMP) is presented passing throw many stages according to the initiated GMP Systematic Structure Stages flowchart in Fig. 5. The 1st stage steps will be as follows;

1. Calculating Gamma matrix of the Breast Cancer Training Data **Table 2**.

2. Classify the values of Gamma. By classifying the Gamma values of the conditional attributes given in the 1st stage to four classes; (G = 1, G = -1, 0 < G <1, -1 < G < 0) and obtaining the conditional attributes classes according this division as in **(Fig. 1. B)**.

3. Compare the conditional attributes row classes of the previous step to get the training data Gamma Classifiers **(Fig. 1.C)**. Comparing the classifiers of with each other, we will find 5 classifiers, 3 (Light Blue Shaded Columns), and the other two (Light Yellow Shaded Columns) are identical.
4. Applying the misclassification error (M_{Error}) rule defined in Equation (3), as a verification method on (Fig.1.C) to get the Breast Cancer Training Data M_{Error} in **(Fig. 1.D).**
5. According to M_{Error} values of the 4th stage with the G classifiers of the 3rd step above, the suitable mining classifiers are G = -1 & -1< G <0 classifiers and they are identical.

Best G Classifier=$\{\{\boldsymbol{A_1}, \boldsymbol{A_{10}}\}, \{\boldsymbol{A_2}, \boldsymbol{A_3}, \boldsymbol{A_5}, \boldsymbol{A_6}, \boldsymbol{A_7}\}, \{\boldsymbol{A_4}\}, \{\boldsymbol{A_8}\}, \{\boldsymbol{A_9}\}\}$ (**5**)

6. The dimension of the above classifier is 5 sets. Hence, we want to reduce the dimension size of our classifier using Gamma average absorption reduction as follows in the 2nd stage.

4. GMP 2ND STAGE: GAMMA MINING PROCEDURE CLASSIFIER

1. Merging the conditional attributes from the G classifiers obtained in the 1st stage; G = -1 & -1< G < 0; **to reduce the dimensional size of the conditional attributes of Gamma Matrix**, then form a new matrix using Gamma Average ($G_{Ave.}$), definition in the methods section. Then, we got the Training Data Gamma Average ($G_{Ave.}$) Matrix in **(Fig. 2.A).**
2. Following the same process of the 2nd step in the 1st stage and applying them on (Fig. 1. B) to get Gamma Average Classes **(Fig. 2. B).**
3. Gamma absorption process for the Gamma average classes (Fig. 2. B) will be calculated defined in the methods section to get **(Fig. 2. C).**
4. Deciding the best $G_{Ave.}$ Absorption classifier of (Fig. 2. C) depends on the conditional attributes M_{Error} values in **(Fig. 1.D).** So, the GMP classifier will be $G_{Ave.}$ = -1 because it is the best absorption $G_{Ave.}$ Classifier as follows.

$$\textit{\textbf{GMP Classifier}} = \{\{\boldsymbol{A_1}, \boldsymbol{A_{10}}\}, \{\boldsymbol{A_2}, \boldsymbol{A_3}, \boldsymbol{A_4}, \boldsymbol{A_5}, \boldsymbol{A_6}, \boldsymbol{A_7}\}, \{\boldsymbol{A_8}, \boldsymbol{A_9}\}\} \quad \textit{\textbf{(6)}}$$

5. GMP 3RD STAGE: GMP PREDICTOR

1. A new predictor using GMP classifier will be initiated in this section depending on calculating M_{Error} average for each class in Relation (6) of the *GMP classifier* and comparing the value of each class in the classifier by the M_{Error} of the diagnosis attribute to get **Fig. (3.A).**
2. According to (Fig. 3.A), **Diagnosis M_{Error} = 0.3**.& the closest M_{Error} average class value to the Diagnosis M_{Error} is the class $\{A_2, A_3, A_4, A_5, A_6, A_7\}. So,$ the GMP predictor attributes as in **(Fig. 3.B)** will be as follows,

$$\textbf{GMP Predictor} = \{\boldsymbol{A_2}, \boldsymbol{A_3}, \boldsymbol{A_4}, \boldsymbol{A_5}, \boldsymbol{A_6}, \boldsymbol{A_7}\} \quad \textbf{(7)}$$

6. GMP 4TH STAGE: TEST DATA NAÏVE BAYES PREDICTION PROCESS

1. In the beginning we will reduce the test data attributes in Table (3) by using the GMP Predictor=$\{A_2, A_3, A_4, A_5, A_6, A_7\}$ as in **(Fig. 4.A).**
2. According to the predictor attributes $\{A_2, A_3, A_4, A_5, A_6, A_7\}$ characteristics of each patient in Table (3) of the Breast Cancer test data, the patients whom have the same values of the predictor attributes can be mined in **(Fig. 4.B).**
3. Using Naïve Bayes as a predictor technique for the test data & using the predictor attributes$\{A_2, A_3, A_4, A_5, A_6, A_7\}$. The differentiation of the predicted diagnosis takes two values, Malignant (D =1) or Benign (D=2). Obtaining Naïve Bayes prediction

calculations will be obtained by applying Relation (4) on (Fig. 4.B). Then the predicted diagnosis of the test data demonstrated in **(Fig. 4.C).**

7. GMP 5THSTAGE: GMP FLOW CHART

In this part **the flow chart** of the obtained GMP is presented as in **(Fig. 5)** clarifying all the above steps by the same systematic order.

8. GMP 6THSTAGE: GMP VIA RAPID MINER VALIDATION

In this section we will validate the GMP technique using **Rapid Miner** because it has the highest rank to use between data mining softwares [7,10,14]. The applicable breast cancer data are analyzed applying K-Nearest Neighbor (KNN) classification model using bootstrapping validation for the classification of the training & test data. **The predicted result was 100% identical with the new GMP technique**. The steps of this part are introduced in **(Fig. 6**).

9. GMP 7THSTAGE: GMP VIA SPSS MODELER VALIDATION

Another validation step of our new GMP using **SPSS Modeler** [10,14] of the cluster analysis of the breast cancer data using the Two Step cluster analysis technique because it has the highest quality. **The result was 99.83 % matching with GMP result**. All the steps of this part are presented in **(Fig. 7)**.

10. TABLES AND FIGURES

Table 1. Breast Cancer Data Description

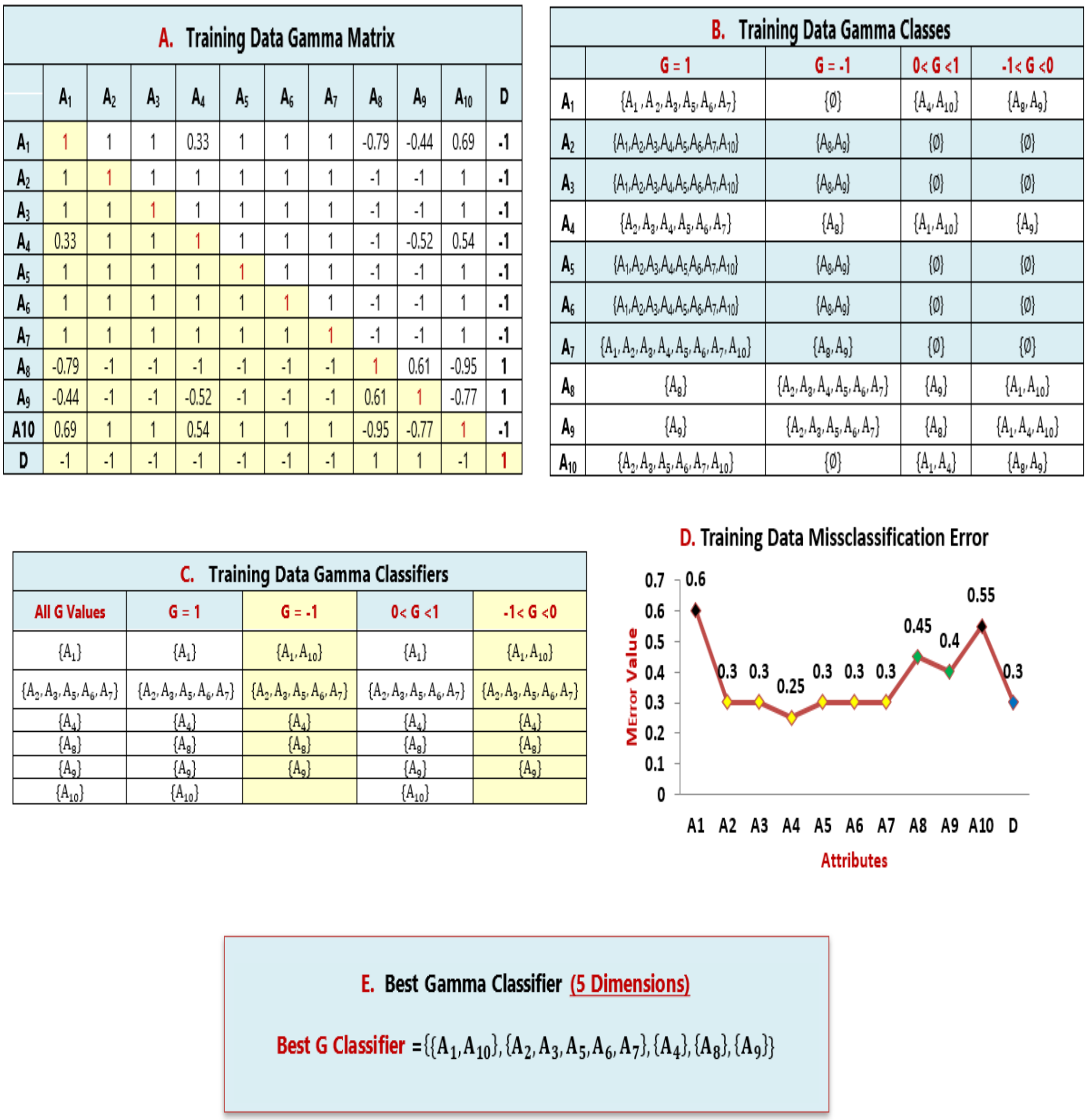
At. Symbol	At. Description	At. Division	At. Code
A_1	Mitosis	No mitosis	1
		1-3 Mitosis	2
		4-7 Mitosis	3
		≥ 8 → Mitosis	4
A_2	Chromatin	Normal Chromatin	1
		Hyper Chromatin	2
A_3	N/C Ratio	Normal	1
		High	2
A_4	NecrosisCellular	Absent	1
		Minimal	2
		Moderate	3
		Marked	4
A_5	Myoepithelial Cells	Negative (-ve)	1
		Positive (+ve)	2
A_6	Basement Membrane	Negative (-ve)	1
		Positive (+ve)	2
A_7	Nuclear Shape	Irregular	1
		Regular	2
A_8	Desmoplasia	Negative (-ve)	1
		Positive (+ve)	2
A_9	Nodal Metastasis	Negative (-ve)	1
		Positive (+ve)	2
A_{10}	Grading	No Grading	1
		Grade I	2
		Grade II	3
		Grade III	4
D	Diagnosis	Malignant	1
		Benign	2

Table 2. Breast Cancer Training Data

At. No. Patient No.	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	D
P_1	4	2	2	1	2	2	2	1	1	4	1
P_2	3	2	2	1	2	2	2	1	1	3	1
P_3	2	2	2	2	2	2	2	1	2	3	1
P_4	4	2	2	1	2	2	2	2	2	1	1
P_5	2	2	2	2	2	2	2	1	1	3	1
P_6	1	1	1	1	1	1	1	2	2	1	2
P_7	1	1	1	1	1	1	1	2	2	1	2
P_8	3	2	2	3	2	2	2	1	1	3	1
P_9	3	2	2	2	2	2	2	1	1	3	1
P_{10}	1	1	1	1	1	1	1	2	2	1	2
P_{11}	1	1	1	1	1	1	1	2	2	1	2
P_{12}	3	2	2	1	2	2	2	1	2	3	1
P_{13}	3	2	2	1	2	2	2	1	2	3	1
P_{14}	3	2	2	1	2	2	2	1	2	3	1
P_{15}	3	2	2	2	2	2	2	1	2	1	1
P_{16}	1	1	1	1	1	1	1	2	2	1	2
P_{17}	1	1	1	1	1	1	1	2	2	1	2
P_{18}	2	2	2	1	2	2	2	2	1	3	1
P_{19}	1	2	2	1	2	2	2	2	1	1	1
P_{20}	3	2	2	1	2	2	2	1	1	2	1

Table 3. Breast Cancer Test Data

~~At. No. Patient No.~~	~~A_1~~	~~A_2~~	~~A_3~~	~~A_4~~	~~A_5~~	~~A_6~~	~~A_7~~	~~A_8~~	~~A_9~~	~~A_{10}~~
~~P_1~~	2	2	2	2	2	2	2	1	1	3
~~P_2~~	1	1	1	1	1	1	1	2	2	1
~~P_3~~	1	2	2	2	2	2	2	1	1	1
~~P_4~~	3	2	2	2	2	2	2	1	1	3
~~P_5~~	3	2	2	3	2	2	2	2	1	3
~~P_6~~	3	2	2	1	2	2	2	1	1	3
~~P_7~~	3	2	2	3	2	2	2	1	1	4
~~P_8~~	2	2	2	1	2	2	2	1	1	1
~~P_9~~	1	1	1	1	1	1	1	2	2	1
~~P_{10}~~	1	1	1	1	1	1	1	2	2	1



A. Training Data Gamma Matrix

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	D
A_1	1	1	1	0.33	1	1	1	-0.79	-0.44	0.69	-1
A_2	1	1	1	1	1	1	1	-1	-1	1	-1
A_3	1	1	1	1	1	1	1	-1	-1	1	-1
A_4	0.33	1	1	1	1	1	1	-1	-0.52	0.54	-1
A_5	1	1	1	1	1	1	1	-1	-1	1	-1
A_6	1	1	1	1	1	1	1	-1	-1	1	-1
A_7	1	1	1	1	1	1	1	-1	-1	1	-1
A_8	-0.79	-1	-1	-1	-1	-1	-1	1	0.61	-0.95	1
A_9	-0.44	-1	-1	-0.52	-1	-1	-1	0.61	1	-0.77	1
A10	0.69	1	1	0.54	1	1	1	-0.95	-0.77	1	-1
D	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1

B. Training Data Gamma Classes

	G = 1	G = -1	0< G <1	-1< G <0
A_1	$\{A_1, A_2, A_3, A_5, A_6, A_7\}$	$\{\emptyset\}$	$\{A_4, A_{10}\}$	$\{A_8, A_9\}$
A_2	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_3	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_4	$\{A_2, A_3, A_4, A_5, A_6, A_7\}$	$\{A_8\}$	$\{A_1, A_{10}\}$	$\{A_9\}$
A_5	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_6	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_7	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_8	$\{A_8\}$	$\{A_2, A_3, A_4, A_5, A_6, A_7\}$	$\{A_9\}$	$\{A_1, A_{10}\}$
A_9	$\{A_9\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_8\}$	$\{A_1, A_4, A_{10}\}$
A_{10}	$\{A_2, A_3, A_5, A_6, A_7, A_{10}\}$	$\{\emptyset\}$	$\{A_1, A_4\}$	$\{A_8, A_9\}$

C. Training Data Gamma Classifiers

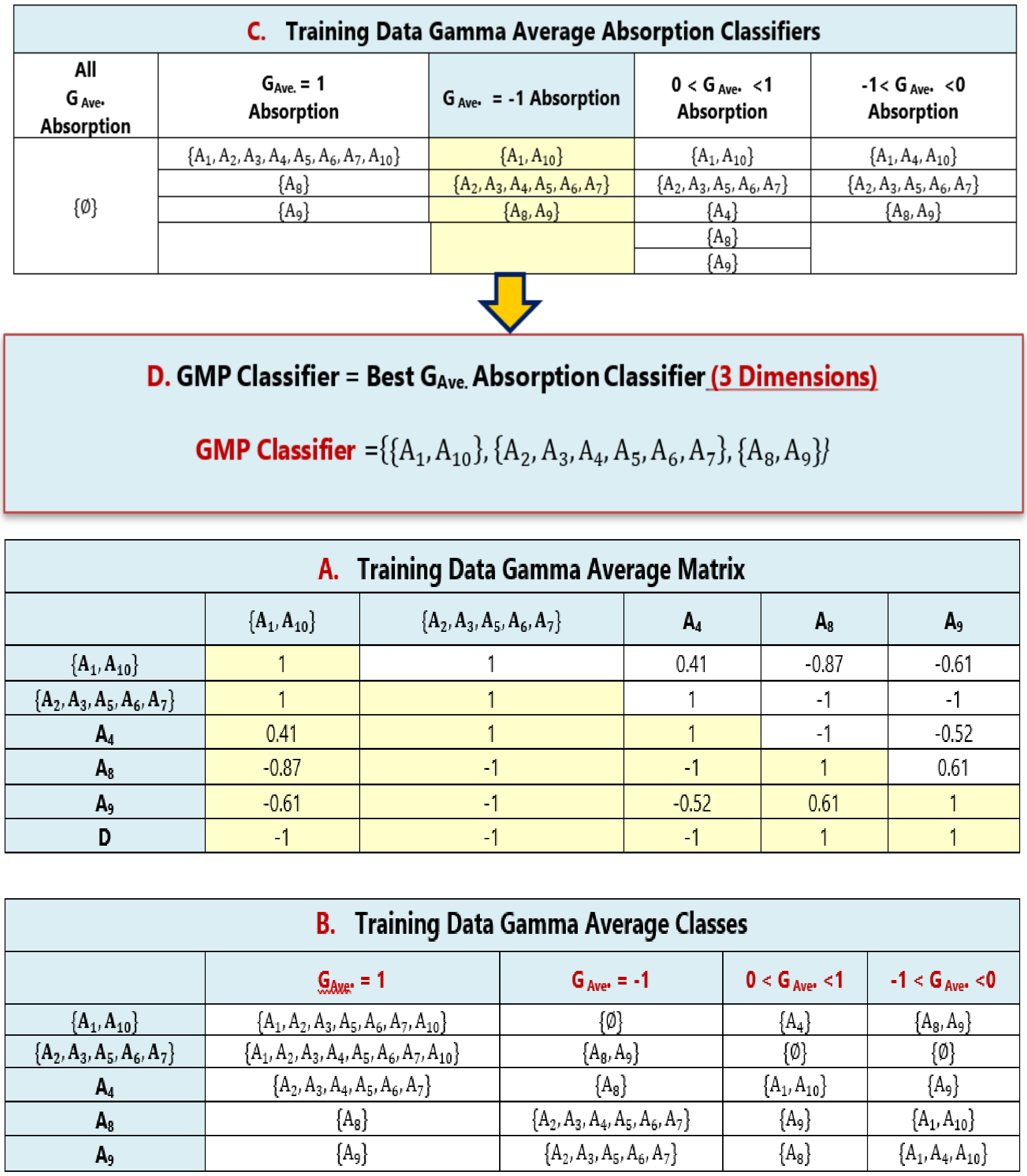
All G Values	G = 1	G = -1	0< G <1	-1< G <0
$\{A_1\}$	$\{A_1\}$	$\{A_1, A_{10}\}$	$\{A_1\}$	$\{A_1, A_{10}\}$
$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$
$\{A_4\}$	$\{A_4\}$	$\{A_4\}$	$\{A_4\}$	$\{A_4\}$
$\{A_8\}$	$\{A_8\}$	$\{A_8\}$	$\{A_8\}$	$\{A_8\}$
$\{A_9\}$	$\{A_9\}$	$\{A_9\}$	$\{A_9\}$	$\{A_9\}$
$\{A_{10}\}$	$\{A_{10}\}$		$\{A_{10}\}$	

D. Training Data Missclassification Error

E. Best Gamma Classifier (5 Dimensions)

Best G Classifier $=\{\{A_1, A_{10}\}, \{A_2, A_3, A_5, A_6, A_7\}, \{A_4\}, \{A_8\}, \{A_9\}\}$

Figure 1. GMP 1st Stage: Best Gamma Classifier



C. Training Data Gamma Average Absorption Classifiers

All $G_{Ave.}$ Absorption	$G_{Ave.}$ = 1 Absorption	$G_{Ave.}$ = -1 Absorption	0 < $G_{Ave.}$ <1 Absorption	-1< $G_{Ave.}$ <0 Absorption
$\{\emptyset\}$	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_1, A_{10}\}$	$\{A_1, A_{10}\}$	$\{A_1, A_4, A_{10}\}$
	$\{A_8\}$	$\{A_2, A_3, A_4, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_2, A_3, A_5, A_6, A_7\}$
	$\{A_9\}$	$\{A_8, A_9\}$	$\{A_4\}$	$\{A_8, A_9\}$
			$\{A_8\}$	
			$\{A_9\}$	

D. GMP Classifier = Best $G_{Ave.}$ Absorption Classifier (3 Dimensions)

GMP Classifier $=\{\{A_1, A_{10}\}, \{A_2, A_3, A_4, A_5, A_6, A_7\}, \{A_8, A_9\}\}$

A. Training Data Gamma Average Matrix

	$\{A_1, A_{10}\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	A_4	A_8	A_9
$\{A_1, A_{10}\}$	1	1	0.41	-0.87	-0.61
$\{A_2, A_3, A_5, A_6, A_7\}$	1	1	1	-1	-1
A_4	0.41	1	1	-1	-0.52
A_8	-0.87	-1	-1	1	0.61
A_9	-0.61	-1	-0.52	0.61	1
D	-1	-1	-1	1	1

B. Training Data Gamma Average Classes

	$G_{Ave.}$ = 1	$G_{Ave.}$ = -1	0 < $G_{Ave.}$ <1	-1 < $G_{Ave.}$ <0
$\{A_1, A_{10}\}$	$\{A_1, A_2, A_3, A_5, A_6, A_7, A_{10}\}$	$\{\emptyset\}$	$\{A_4\}$	$\{A_8, A_9\}$
$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{10}\}$	$\{A_8, A_9\}$	$\{\emptyset\}$	$\{\emptyset\}$
A_4	$\{A_2, A_3, A_4, A_5, A_6, A_7\}$	$\{A_8\}$	$\{A_1, A_{10}\}$	$\{A_9\}$
A_8	$\{A_8\}$	$\{A_2, A_3, A_4, A_5, A_6, A_7\}$	$\{A_9\}$	$\{A_1, A_{10}\}$
A_9	$\{A_9\}$	$\{A_2, A_3, A_5, A_6, A_7\}$	$\{A_8\}$	$\{A_1, A_4, A_{10}\}$

Figure 2. GMP 2nd Stage: Gamma Mining Procedure Classifier

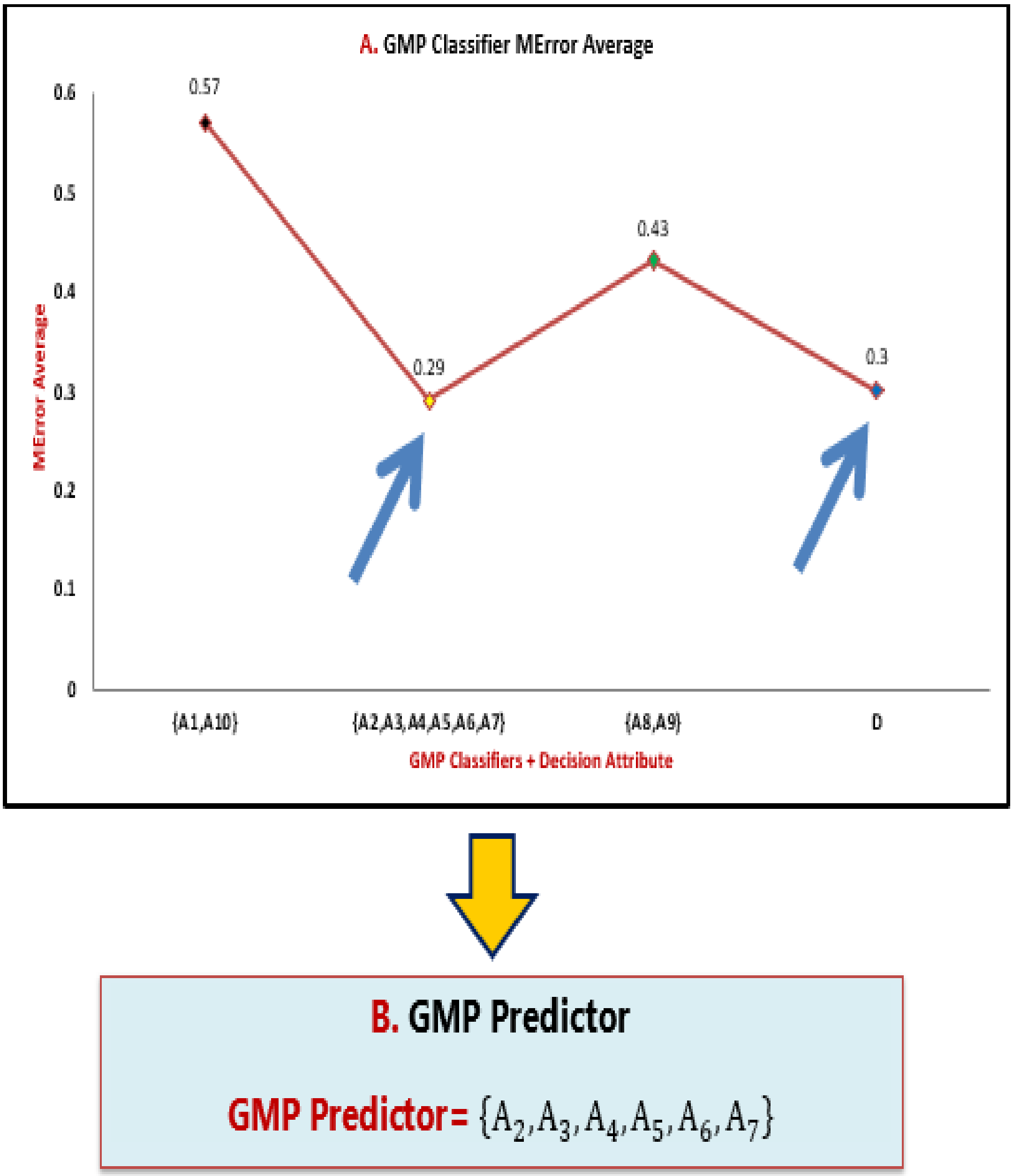
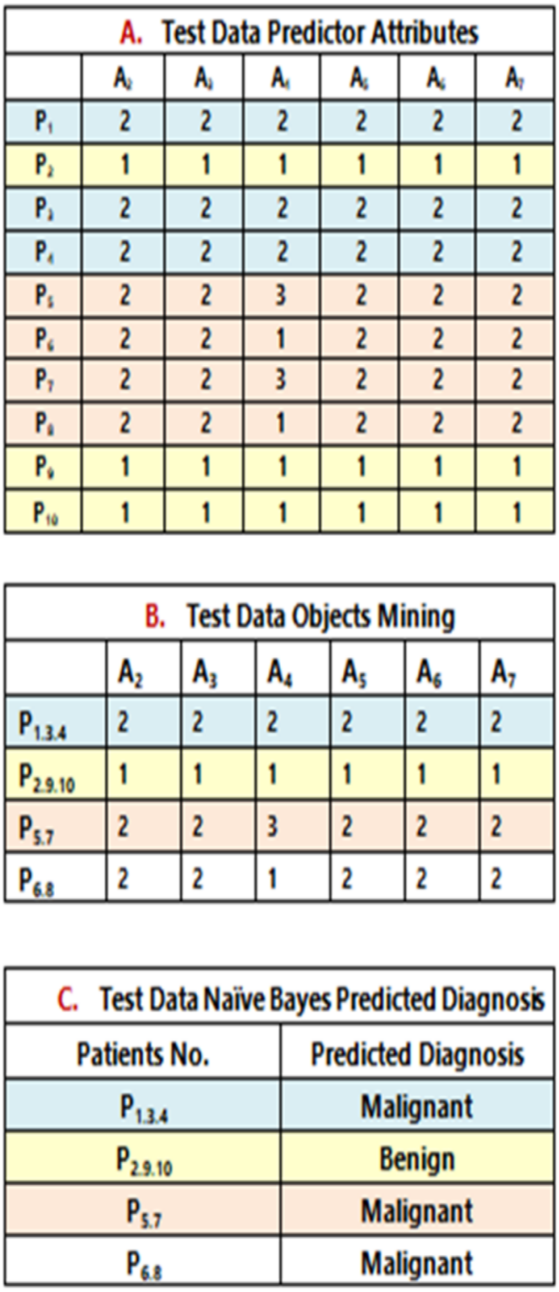


Figure 3. GMP 3rd Stage: GMP Predictor



A. Test Data Predictor Attributes

	A_2	A_3	A_4	A_5	A_6	A_7
P_1	2	2	2	2	2	2
P_2	1	1	1	1	1	1
P_3	2	2	2	2	2	2
P_4	2	2	2	2	2	2
P_5	2	2	3	2	2	2
P_6	2	2	1	2	2	2
P_7	2	2	3	2	2	2
P_8	2	2	1	2	2	2
P_9	1	1	1	1	1	1
P_{10}	1	1	1	1	1	1

B. Test Data Objects Mining

	A_2	A_3	A_4	A_5	A_6	A_7
$P_{1.3.4}$	2	2	2	2	2	2
$P_{2.9.10}$	1	1	1	1	1	1
$P_{5.7}$	2	2	3	2	2	2
$P_{6.8}$	2	2	1	2	2	2

C. Test Data Naïve Bayes Predicted Diagnosis

Patients No.	Predicted Diagnosis
$P_{1.3.4}$	Malignant
$P_{2.9.10}$	Benign
$P_{5.7}$	Malignant
$P_{6.8}$	Malignant

Figure 4. GMP 4th Stage: Test Data Naïve Bayes Prediction Process

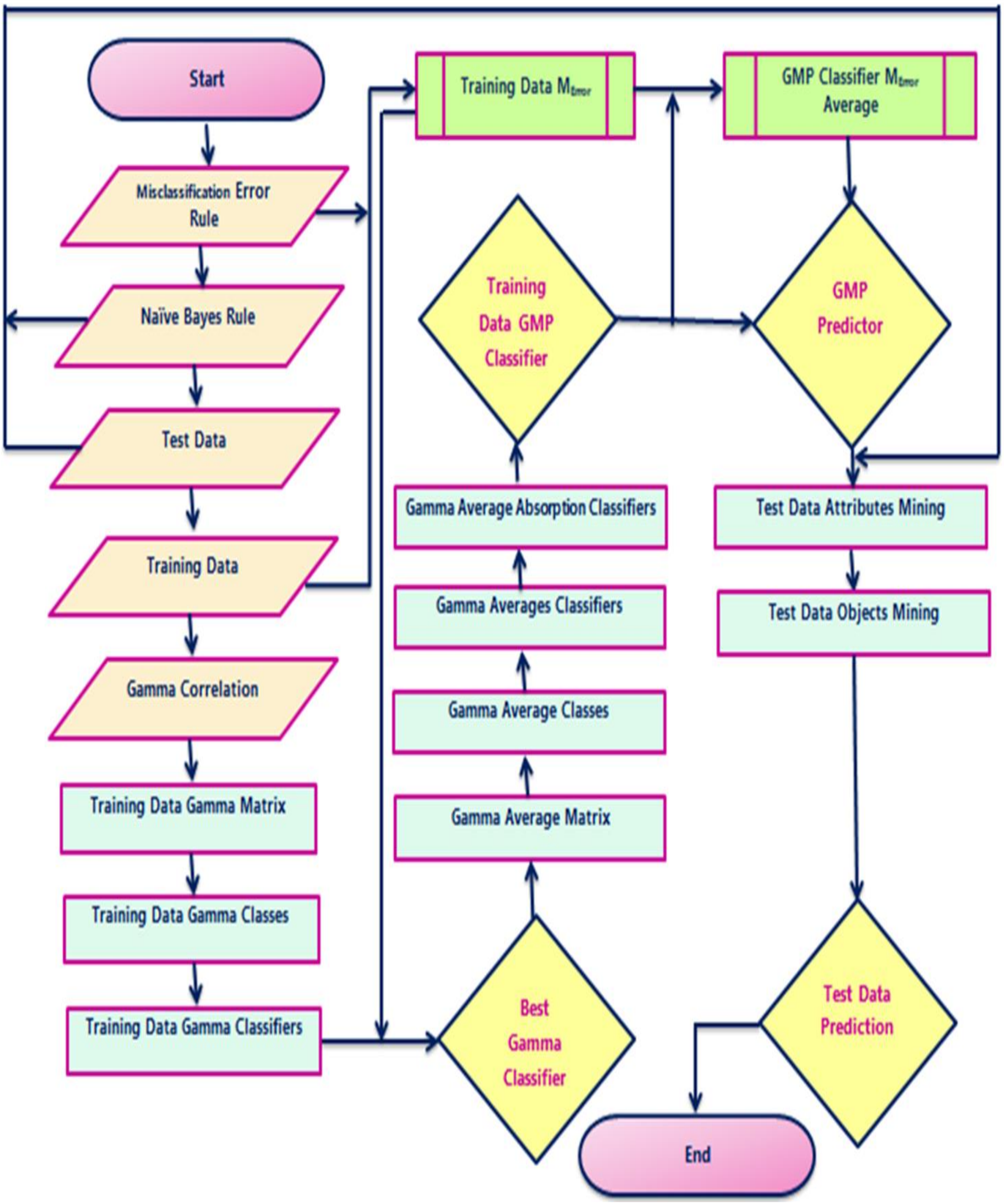
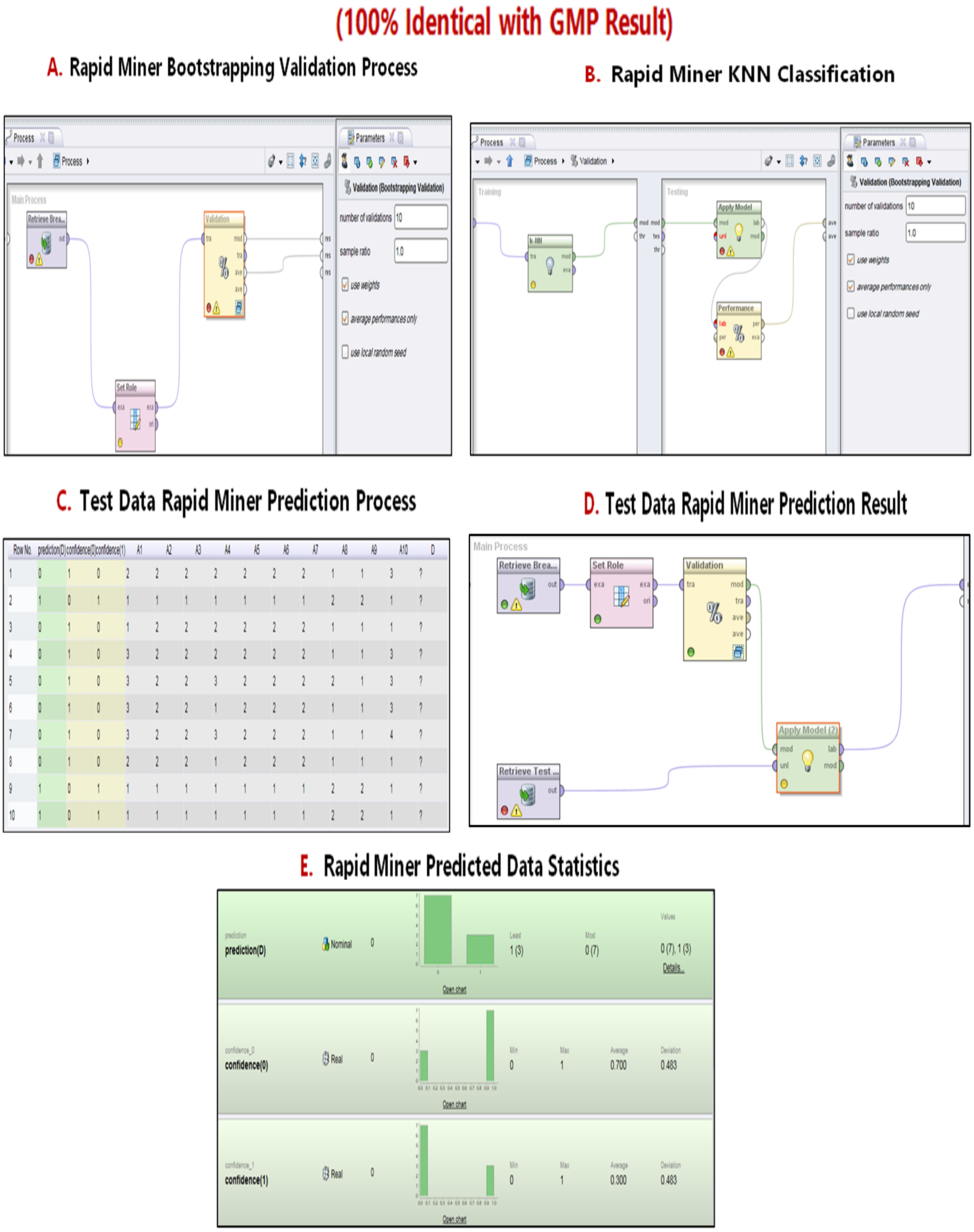

Start
Misclassification Error Rule
Naïve Bayes Rule
Test Data
Training Data
Gamma Correlation
Training Data Gamma Matrix
Training Data Gamma Classes
Training Data Gamma Classifiers
Training Data M_{Error}
GMP Classifier M_{Error} Average
Training Data GMP Classifier
GMP Predictor
Gamma Average Absorption Classifiers
Gamma Averages Classifiers
Gamma Average Classes
Gamma Average Matrix
Best Gamma Classifier
Test Data Attributes Mining
Test Data Objects Mining
Test Data Prediction
End

Figure 5. GMP 5th Stage: GMP Flow Chart



(100% Identical with GMP Result)

A. Rapid Miner Bootstrapping Validation Process

Validation (Bootstrapping Validation)

number of validations 10

sample ratio 1.0

☑ use weights

☑ average performances only

□ use local random seed

B. Rapid Miner KNN Classification

Validation (Bootstrapping Validation)

number of validations 10

sample ratio 1.0

☑ use weights

☑ average performances only

□ use local random seed

C. Test Data Rapid Miner Prediction Process

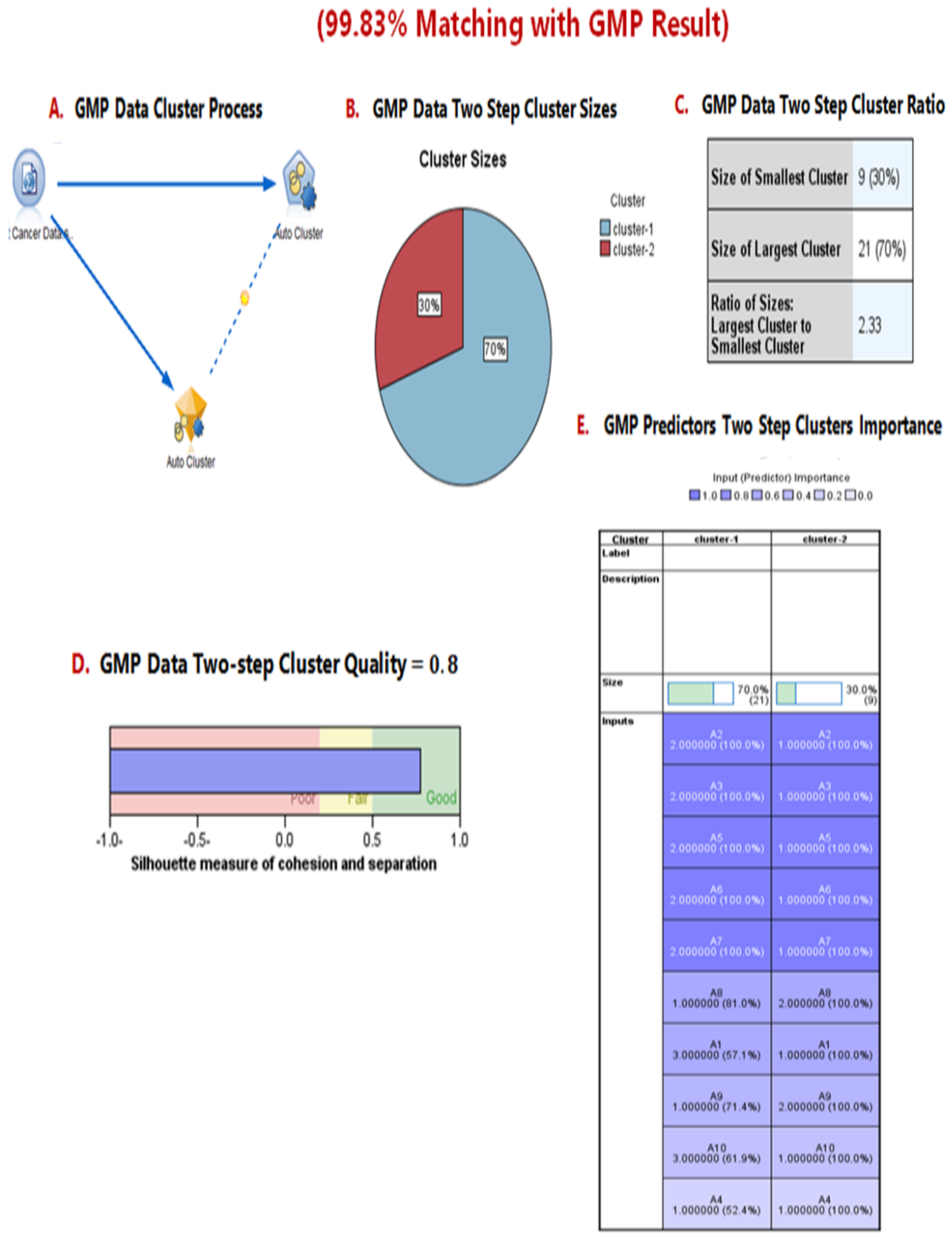
Row No.	prediction(D)	confidence(0)	confidence(1)	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	D
1	0	1	0	2	2	2	2	2	2	2	1	1	3	?
2	1	0	1	1	1	1	1	1	1	1	2	2	1	?
3	0	1	0	1	2	2	2	2	2	2	1	1	1	?
4	0	1	0	3	2	2	2	2	2	2	1	1	3	?
5	0	1	0	3	2	2	3	2	2	2	2	1	3	?
6	0	1	0	3	2	2	1	2	2	2	1	1	3	?
7	0	1	0	3	2	2	3	2	2	2	1	1	4	?
8	0	1	0	2	2	2	1	2	2	2	1	1	1	?
9	1	0	1	1	1	1	1	1	1	1	2	2	1	?
10	1	0	1	1	1	1	1	1	1	1	2	2	1	?

D. Test Data Rapid Miner Prediction Result

E. Rapid Miner Predicted Data Statistics

prediction prediction(D)	Nominal	0	Least 1 (3)	Most 0 (7)		Values 0 (7), 1 (3) Details...
confidence_0 confidence(0)	Real	0	Min 0	Max 1	Average 0.700	Deviation 0.483
confidence_1 confidence(1)	Real	0	Min 0	Max 1	Average 0.300	Deviation 0.483

Figure 6. GMP 6th Stage: GMP via Rapid Miner Validation



A. GMP Data Cluster Process

B. GMP Data Two Step Cluster Sizes

C. GMP Data Two Step Cluster Ratio

Size of Smallest Cluster	9 (30%)
Size of Largest Cluster	21 (70%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.33

D. GMP Data Two-step Cluster Quality = 0.8

E. GMP Predictors Two Step Clusters Importance

Cluster	cluster-1	cluster-2
Label		
Description		
Size	70.0% (21)	30.0% (9)
Inputs	A2 2.000000 (100.0%)	A2 1.000000 (100.0%)
	A3 2.000000 (100.0%)	A3 1.000000 (100.0%)
	A5 2.000000 (100.0%)	A5 1.000000 (100.0%)
	A6 2.000000 (100.0%)	A6 1.000000 (100.0%)
	A7 2.000000 (100.0%)	A7 1.000000 (100.0%)
	A8 1.000000 (81.0%)	A8 2.000000 (100.0%)
	A1 3.000000 (57.1%)	A1 1.000000 (100.0%)
	A9 1.000000 (71.4%)	A9 2.000000 (100.0%)
	A10 3.000000 (61.9%)	A10 1.000000 (100.0%)
	A4 1.000000 (52.4%)	A4 1.000000 (100.0%)

Figure 7. GMP 7th Stage: GMP via SPSS Modeler Validation

11. CONCLUSION & FUTURE WORK

A new Statistical Data Mining (SDM) technique is initiated in this paper using Gamma Mining Procedure (GMP) contributing a novel classifier & predictor depending on Gamma (G) correlation coefficient matrix, Misclassification Error (M_{Error}) is used as an accuracy measure and Naive Bayes as a prediction technique. Applying GMP on real life application of Breast Cancer tumor diagnosis and use it for predicting the diagnosis of the test data to see the accuracy of its applicability and the result was typical for the diagnosis of the data specialist. GMP opens the way for other new SDM techniques using an alternative correlation coefficient according to the data type.

The New GMP validation is obtained by using the most two applicable soft wares of the SDM process which are (Rapid Miner & SPSS Modeler) and the result was 100% identical with Rapid Miner result, while, the GMP result was 99.83% matching with SPSS Modeler Analysis. These strong validations leads to that GMP is a very strong effective technique to use it in the SDM process.

REFERENCES

[1] Abd El-Badie S. (2006) “A New Data Reduction Approach”, MS. C. Thesis, Faculty of Science, Tanta University, Egypt.

[2] Abd El-Badie S. (2014) “Statistics & Data Mining”, Ph. D. Thesis, Faculty of Science, Tanta University, Egypt.

[3] Abd El-Monsef M.E., Rady E. A., Kozea A. M., Hassanein W. A. and Abd El-Badie S. (2013) “What is the Major Power Linking Statistics & Data Mining?”, International Journal of Data Mining & Knowledge Management Process. Volume 3, Number 6, ISSN 2230-9608, AIRCC Co.

[4] Abd El-Monsef M.E., Rady E. A., Kozea A. M., Hassanein W. A. and Abd El-Badie S. (2014) “A New SDM Classifier Using Jaccard Mining Procedure Case Study: Rheumatic Fever Data “International Journal on Bioinformatics & Biosciences (IJBB) Vol. 4. Issue (1).

[5] Annabel Ross (2020) “Ordinal Measures of Correlations” Chapter 14. Web: [PDF] CHAPTER 14 ORDINAL MEASURES OF CORRELATION: SPEARMAN\\\’S RHO AND GAMMA – Free Download PDF (silo.tips)

[6] Brett Balloun (2013) “Complement for Microsoft Basic Data Mining Tutorial”. Web: Complement for Microsoft Basic Data Mining Tutorial - Perficient Blogs

[7] Galit Shmueli, Peter C. Bruce, Amit V. Deokar & Nitin R. Patel. (2022) “Machine Learning for Business Analytics: Concepts, Techniques and Applications in Rapid Miner”, 1sted.ISBN-13: 978-1119828792.

[8] Han J. and Kamber M. (2006) “Data Mining: Concepts and Technique”, 2nd ed., Editor Morgan Kaufmann Publishers,ISBN 1-55860-901-6.

[9] Healey J. F. 92010) “STATISTICS: A Tool for Social Research”, 9th ed. Wadsworth USA.

[10] IBM Corp. Released (2011), IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

[11] Machine Learning Summary Notes (2021) Web: CZ4041 Machine Learning Summary Notes | CZ4041 - Machine Learning - NTU | Thinkswap

[12] Microsoft Doc. (2022) “Training and Testing Data Sets “Article. Web: Training and Testing Data Sets Microsoft Docs

[13] Shubhnoor Gill. (2021) “12 Top Data Mining Tools in 2022”. December 21st. Web: 12 Top Data Mining Tools in 2022 - Learn | Hevo (hevodata.com)

[14] SPSS Inc. (2005) “SPSS Data Mining Tips”, ISBN 1-56827-282-0 Printed in the U.S.A.

[15] StatSoft Inc. (2013) “Electronic Version: Electronic Statistics Textbook. Tulsa”, StatSoft. Web: http://www.statsoft.com/textbook/naive-bayes-classifier

[16] Tan, Steinbach M., Kumar V. (2005) “Introduction to Data Mining”, Addison-Wesley.

[17] Wikipedia (2020) “Goodman and Kruskal Gamma” Article. Web: Goodman and Kruskal's gamma - Wikipedia

AUTHOR

Dr. Soaad Abd El-Badie Attia El-Afefy (S. Abd El-Badie) , Ph. D., Titled “Statistics and Data Mining”, 2014, Mathematics Department, Tanta University, Egypt. She got her MS. C. Titled ”A New Data Reduction Approach”, 2006, Faculty of Science, Tanta University, Egypt. She got the Best Student Presentation and Best Student Paper Awards in many conferences. She worked in the academic career for more than ten years in many places (GUC Mathematic Department, Tanta University Faculty of Science, Kafr El-Sheikh University Faculty of Engineering, CAPMAS, and Cairo Academy) and Now, She is working in The Information Technology Sector, Man Power & Immigration Directorate, Tanta, Egypt. Homepage: http://www.savvymore.mysite.com