VOL. 12, NO. 11, JUNE 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MODEL OF GENETIC FUZZY ARTMAP CLASSIFIER (GFAM) FOR GASTRIC CANCER DATA CLASSIFICATION

Thara Lakshmipathy[1] and Gunasundari Ranganathan[2]
[1]Department of Computer Science, Karpagam Academy of Higher Education, Tamil Nadu, India
[2]Department of Information Technology, Karpagam University, Tamil Nadu, India
E-Mail: kltharavijay@gmail.com

ABSTRACT

Data mining is the evergreen research area in the field of Computer Science. Its artefact applies in the area of healthcare, decision support and expert systems. Soft computing plays a significant role in the design and development of predictive and descriptive data mining applications. This research work presents a fuzzy adaptive resonance theory classifier with the support of genetic algorithm for gastric cancer data classification. The metrics such as accuracy, hit rate and elapsed run time are chosen for the performance evaluation. From the results it is evident that the GFAM attains better performance.

Keywords: gastric cancer data classification, data mining, fuzzy logic, Q-learning, genetic algorithm.

1. INTRODUCTION

Data mining clutches remarkable scope for the healthcare sector in order to facilitate health systems and health professionals to scientifically make use of the data and analytics for recognizing the comforts and best practices which leads to elevated patient care and also results in lessening the costs. A survey on Data Mining Techniques that have been employed for Bio Medical Research, presents the significance of such algorithms in the disease diagnosing process [14].

As per the statistics on the death rate assessment, cancer becomes the foremost cause of death even in the developed nations and stands the second major cause of death in the developing nations. Particularly, gastric cancer stands the fourth-most common cancer and stands at the second leading cause of cancer deaths at the international level. This paves the motivation to auxiliary inspect the issues distressing the occurrence of the disease owing to the pervasiveness of the disease and the high mortality rate of gastric cancer. This research work aims to design and develop a model for gastric cancer data classification. At the initial stage Fuzzy ARTMAP (FAM) classifier with Q-learning (known as QFAM) is developed for incremental learning of data samples. Next it is aimed to make use of genetic algorithm (GA) for the rule extraction from QFAM.

2. BACKGROUND

Certain techniques have effectively been recognized to decide, separate, and order subtypes of gastric cancer (GC) and to comprehend some symptomatic predicaments [1]. The most common factor that leads to stomach cancer is H. Poly which is the most considerable threat in the human body. The cancer cells start creating from the inner region of the stomach. If these cells are allowed to grow, they could form a tumour which could spread slowly inside the stomach over years together to form Gastric cancer [15]. In early gastric cancer (EGC), tumour attack is limited to the mucosa or sub mucosa paying little mind to the nearness of the lymph hub metastasis or not [2]. Gene expression investigation recognizes a signature that separated EGC from ordinary tissue [2]. Boussioutas *et al*. break down 124 tumour and contiguous mucosa tests and investigate the sub-atomic elements of gastric cancer, which could be perceived that promptly characterizes the premalignant and tumour subtypes, utilizing DNA microarray-based gene expression profiling [3]. The recognizable proof of the atomic signatures that are normal for the subtypes of gastric cancer and related premalignant changes ought to empower the assist examination of the means required in the start and movement of gastric cancer. Vecchi *et al.* inferred 1024 genes (52% up-controlled and 48% down-directed) that are differentially communicated in 19 EGC tests when contrasted with 9 typical tissues [4].

The up-directed genes are included in cell cycle, RNA handling, ribosome biogenesis, and cytoskeleton association, while the down-control genes are embroiled in particular elements of the gastric mucosa (assimilation, lipid digestion system, and G-protein-coupled receptor protein flagging pathway). Nam *et al*. [5] likewise distinguished a 973-gene signature to separate the EGC from the ordinary tissue utilizing the microarray information from the coordinated tumour and neighbouring non-cancerous tissues of 27 EGC patients [5]. They promote showed that the up-directed genes in EGC tissues are related with cell relocation and metastasis. Kim *et al.* exhibit that 60 genes are continuously up or down-managed in progression in typical mucosa, adenoma, and carcinoma tests by looking at the expression profiles of these tissues from eight patient-coordinated sets. Therefore, atomic order appears to be exceptionally encouraging for sub-atomic analysis of EGC [6].

Both unending gastritis (ChG) and intestinal metaplasia (IM) are included in the middle of the road phase of GC, the previous is portrayed by a mitochondria-related gene expression signature while the last is described by the markers of multiplication. Since ChG has mitochondria gene expression signature, it may enthusiasm to test whether such a signature is identified with the metabolic subtype signature of GC [7]. For sure, the differential communicated gene is set amongst the

ChG and IM, is to a great extent covered with the GC metabolic signature (P = 0.00085, hyper geometric test).

Cancer of unknown primary site (CUP) is very much perceived clinical issue, representing 3-5% of all the dangerous epithelial tumours. Glass can be recognized in the light of preserved tissue, particular in gene expression [8]. It has been demonstrated that gene expression profiling can distinguish tissue starting point with an exactness rate somewhere around 33% and 93% [9]. Anthony et al. connect a 92-gene CUP to examine tumour tests from patients with CUP. Fifteen of 20 cases (75%) are effectively anticipated, i.e., those anticipated CUPs are the genuine inactive essential locales that are recognized after the underlying finding of CUP. This measure has been effectively connected to numerous cancers, for example, bosom, colorectal, and melanoma [10].

These gene signature-based techniques can likewise be utilized to distinguish the particular treatment for GC patients, i.e., focused on treatments. In a substantial planned trial (n = 289), a gene expression signature is created to anticipate the tissue of starting point in many patients with CUP. The middle survival time is 12.5 months for patients who got coordinated site-particular treatment contrasted and the utilization of empiric CUP regimens. Patients whose CUP destinations are anticipated to have more responsive tumour sorts survived longer than those anticipated to have less responsive tumour sorts [11].

3. PRELIMINARIES OF FUZZY ARTMAP (FAM)

FAM is a supervised neural network that has influenced on the incremental learning and is one of the famous ART-based models to resolve classification problems. Certain works are done to augment the performance of FAM, and to relate it to a diversity of data mining applications particularly in healthcare. Feature selection is a procedure of spotting a subset of features obtained through the dataset. Usually, black-box (or pedagogical) and decomposition schemes are employed for performing rule extraction from artificial neural networks. Maximizing the rate of correctly classified patterns and minimizing the number of selected rules are two main challenges.

FAM contains two unsupervised fuzzy ART networks, i.e., ART_a and ART_b, and a map field. The fuzzy ART network consists of three layers. The first layer is the pre-processing layer fa_0^a (fb_0^a). Here complement coding is employed to preprocess the input sample which stays away from the problem of category propagation. The second and third layers are the input layer fa_1^a (fb_1^a) and the recognition layer fa_2^a (fb_2^a), respectively. There are three important parameters for each fuzzy ART network, i.e., the choice parameter mentioned as $\alpha > 0$, vigilance parameter denoted as $\rho \in [0, 1]$, and learning rate represented as $\beta \in [0, 1]$.

At the learning stage, input vector A and its corresponding target vector B are presented to ART_a and ART_b, respectively. The choice function, T_j, is used to measure the similarity between the input pattern and the prototype pattern contained in the weight vector of each node, denoted as j, i.e.

$$T_j = \frac{|A \wedge w_j^a|}{\alpha + |w_j^a|} \quad (1)$$

where $w_j^a \equiv (w_{j1}^a, ..., w_{j2M}^a)$ is the weight vector of the jth node in fa_2^a. Initially, $w_{j1}^a(0) = ... = w_{j2M}^{\alpha}(0) = 1$. The complement coding format is mentioned as follows:

$$w_j^a = (u_j, v_j^c) \quad (2)$$

where u_j and v_j are lower and higher vertices of the jth fa_2^a node, respectively. The prototype node with the highest T_j is chosen and considered as the winning node (denoted as node J), i.e.,

$$T_J = \max\{T_j : j = 1,2,...,N\} \quad (3)$$

where N is the number of prototype nodes. If more than one Tj is maximal, the node with the smallest j index is chosen.

Resonance occurs if the current input and the winning node J satisfy the vigilance test, i.e.,

$$T_j = \max\{T_j : j = 1,2,....N\} \quad (4)$$

where ρ_a is vigilance parameter of ART_a. If the vigilance test in (4) is not satisfied, T_J is set to zero, and a new search cycle (known as match-tracking) to choose a new winning node is triggered. This match-tracking cycle continues until the winning node is able to satisfy the vigilance test.

When no such node exists, a new node is created in f_2^a. In order to find out the winning target node in ART_b, the same process takes place in ART_b simultaneously by utilizing the target vector. When the winning nodes in ART_a and ART_b have been identified, the map-field vigilance test is applied, as follows:

$$\frac{|y^b \wedge W_j^{ab}|}{|y^b|} \geq P_{ab} \quad (5)$$

where $\rho ab \in [0, 1]$ is the map-field vigilance parameter, W_j^{ab} is the weight vector from f_2^a to f^{ab}, and y^b is the output vector of f_2^b. Suppose that the winning node of ART_b is K, then

$$y_k^b = \begin{cases} 1 & if \ \ K = K \\ 0 & otherwise \end{cases} \quad (6)$$

When the map-field vigilance test is not satisfied, match-tracking ensues, and the vigilance parameter of ARTa is updated as follows:

$$\rho_a = \frac{|\ A \wedge w_j^a\ |}{|\ A\ |} + \delta \quad (7)$$

where δ is a small positive value.

By itself, a new search cycle with the new ρ_a setting takes place in ART_a in anticipation of a correct prediction which occurs at the map-field. Once the map-field vigilance test is satisfied, this means that the winning prototype in f_2^a makes a correct prediction of the target class. By default, learning proceeds by which ART_a winning prototype is updated to

$$w_j^{a(new)} = \beta_a (A \wedge w_j^{a(new)}) + (1 - \beta_a) w_j^{a(old)} \quad (8)$$

4. GENETIC FUZZY ARTMAP CLASSIFIER (GFAM)

The GFAM model consists of a two-stage data classification and rule extraction model based on a pruned FAM model and the genetic algorithm. The details of the proposed model, which uses the pruned FAM model in the Stage 1 to reduce network complexity and the GA in the Stage 2 to extract explanatory rules, are explained as follows:

The learning phase of GFAM is the same as that of conventional FAM. However, for each prototype node in f_2^a, a Q-value is incorporated.

During the learning phase of GFAM, the selected winner in f_2^a can result in a correct or an incorrect prediction, which leads to learning or match-tracking, respectively (i.e., in accordance with the FAM learning algorithm). Depending on the prediction result, the winning node in f_2^a can be rewarded (during learning) or penalized (during match-tracking), by using a Q-value, which is updated as follows:

$$Q(j)_t = Q(j)_{t-1} + \xi[\ r(j)_t + \gamma vig(j)_t\ ] \quad (9)$$

where $\gamma \in [0, 1]$ is the discount factor, $\xi \in [0, 1]$ is the learning rate, vig(j) is the vigilance value of the winning node of f_2^a, and r(j) is the reinforcement signal defined as:

$$r(j) = \begin{cases} 1 & \text{if learning happens} \\ 0 & \text{if match - tracking happens} \end{cases} \quad (10)$$

All the nodes that are able to satisfy the vigilance test (Eq. (3)) are recognized for the participation in another selection process. Particularly, the strength of the each selected node is determined by:

$$\text{Strength}(j) = \lambda T(j) + (1-\lambda) Q(j) \quad (11)$$

where $\lambda \in [0, 1]$ is a weighting factor. As can be seen in (11), both the choice function (i.e., T(j)) and Q-value function (i.e., Q(j)) are considered, and the node with the highest strength is selected as the final winner to provide a prediction pertaining to the target output of the current input. This differentiates GFAM from FAM in the prediction phase.

Pruning is performed to reduce the size of f_2^a by removing the less informative prototype nodes. To accomplish this objective, f_2^a nodes with Q-values smaller than a threshold are pruned. After removing f_2^a nodes with low Q-values, the remaining nodes are used in the second stage for rule extraction.

To minimize the number of features in each rule, the remaining f_2^a nodes after pruning are used to create "gastric patients positive" prototypes containing the "unaffected gastric patients" feature. The dimension of the prototype nodes in QFAM is the same as that of the input features. When the dimension is high, the number of antecedents in the extracted rules are also high because each dimension is interpreted as an antecedent. As such, the extracted rules become complicated. In addition to partitioning the input space into a number equivalent to linguistic value, the concept of a "unaffected gastric patients" antecedent is introduced.

One of the main issues in fuzzy rule extraction is the "curse of dimensionality", i.e. the search space for the feasible rules is n_d where n is the number of partitions of each dimension, and d is the number of input features. Once pruning is carried out, the remaining nodes are used to create the "gastric patients positive" prototypes. One of the goals for designing a useful classifier is to have a high classification accuracy rate with a concise rule set having a small number of features. The GA is used for this purpose. The GA chromosome, S, is defined as follows.

$$S = \{D_1^1, D_2^1, \ldots, D_d^1, D_2^1, D_2^1, \ldots, D_d^1, \ldots, DD_1^P, D_2^P, \ldots, D_d^P\} \quad (12)$$

where d is the number of features of each prototype, and p is the number of prototypes after pruning. D_d^P is initialized as follows:

$$D_d^P = \begin{cases} 0 & \text{for don' t care features} \\ 1 & \text{for other features} \end{cases} \quad (13)$$

The fitness function of GA is formulated to maximize classification accuracy and minimize the number of input features, as follows:

Maximize $f(s) = W_{NCP} \cdot NCP(s) - W_S|S|$ (14)

where |S| and NCP(s) are the number of features and the number of correctly classified data samples, respectively, WNCP and WS are two positive weights, and $0 < Ws << W_{NCP}$.

4.1 Classification process

Step 1: Initialization: Create a population of values, N_{pop}. The "unaffected gastric patients" antecedent is denoted as '0', while the rest are denoted as '1'.

Step 2: Selection: Choose $N_{pop}/2$ pairs of values from the current population. The selection probability, P(S), of value S in a population ψ is as follows:

$$P(S) = \frac{\cdot\{f(s) - f_{\min}(\Psi)\}}{\sum_{S \in \Psi}\{f(s) - f_{\min}(\Psi)\}} \quad (15)$$

Where

$$fmin(\psi) = \min \left\langle f(S) \mid S \in \Psi \right\rangle \quad (16)$$

Step 3: Crossover: Based on the crossover probability, randomly select a bit position for each chosen pairs for crossover.

Step 4: Mutation: Apply mutation to the selected values generated in the step 3 based on a mutation probability:

$S_r = 1 \rightarrow S_r = -1$ with probability $P_m(1 \rightarrow -1)$
$Sr = -1 \rightarrow Sr = 1$ with $ProbilityP_m(-1 \rightarrow 1)$ (17)

Step 5: Elitism strategy: Randomly select and remove one of the values from generated values, and add the value with the highest fitness value in the previous population into the current one.

Step6: Termination: If termination condition is satisfied stop, otherwise go to step 1.

4.1. Genetic algorithm parameter settings

Population size 50
Number of generations 470 (set based on number of patient records)
Crossover type= typically two point
Crossover rate of 0.6
Mutation types= bit flip
Mutation rate of 0.001

In rule extraction, each prototype is considered as a fuzzy rule. To facilitate linguistic rule extraction, the input features are quantized. The quantization level, Q, determines the number of fuzzy partitions in the quantized level. The interval of [0, 1] is divided into Q partitions and the round-off method is used for quantization:

$$Vq = \frac{(q-1)}{(Q-1)} \quad (18)$$

where q = 1,...,Q. In GFAM, the extracted fuzzy, if-then rules are as follows:

Rule R_j: IFx_{p1} is V_q and …x_{pn} is V_q, THEN x_p is class C_j with Q_value = Q_values_j where j denotes each prototype node after pruning, V_q is the antecedent value, $x_p = (x_{p1},...,x_{pn})$ is an n-dimensional data vector, and Q_values_j is the Q-value of the jth prototype node calculated using Equation (9). These prototypes are applied in the gastric data classification.

5. RESULTS AND DISCUSSIONS

The dataset are collected from the leading cancer care hospital that contains the records of 470 patients, each of which have 29 features. All features are considered as indicators of gastric cancer for a patient, according to medical literature. However, some of them have never been used in data mining based approaches for gastric cancer diagnosis. The features are arranged in four groups: personal characteristics, personal behaviour, systemic features and the stomach. MATLAB tool has been used for implementing the results.

5.1. Performance evaluation

Performance metrics namely accuracy, hit rate and elapsed run time are considered for comparison. The proposed classifier outperforms the existing algorithms such as Apriori algorithm and ontology based Apriori algorithm. As far as accuracy performance evaluation is concerned, true positive, true negative, false positive, false negative are used to compute accuracy value, as described below:

True positive (TP): Gastric cancer patients correctly identified as affected

True negative (TN): Unaffected patients correctly identified as unaffected

False positive (FP): Unaffected patients incorrectly identified as affected

False negative (FN): Gastric cancer patients incorrectly identified as unaffected

Table-1 calculates the accuracy by adding TP and TN and dividing the result by 470.

Table-1. Performance evaluation on accuracy.

Algorithm	True positive	True negative	False positive	False negative	Accuracy
Apriori Algorithm [13]	291	26	123	30	67.44 %
Ontology based Apriori Algorithm [12]	330	15	54	71	73.40 %
GFAM (Proposed Work)	388	30	19	33	88.95 %

Table-2. Performance Evaluation on Hit Rate and Elapsed Run Time.

Algorithms	Hit rate	Elapsed run time
Apriori Algorithm	64%	2090 seconds
Ontology based Apriori Algorithm	71%	125 seconds
Proposed Classifier	85%	95 seconds

5.2. Result graphs and inference

The Figure-1 depicts the quantitative analysis of the performance evaluation accuracy, shown in the Table-1 as:
(a)True Positive, (b) True Negative, (c) False Positive and (d) False Negative. The Figure-2 depicts the performance analysis of accuracy of the existing algorithms namely Apriori algorithm, Ontology based Apriori algorithm and the proposed algorithm GFAM using the Table-1.

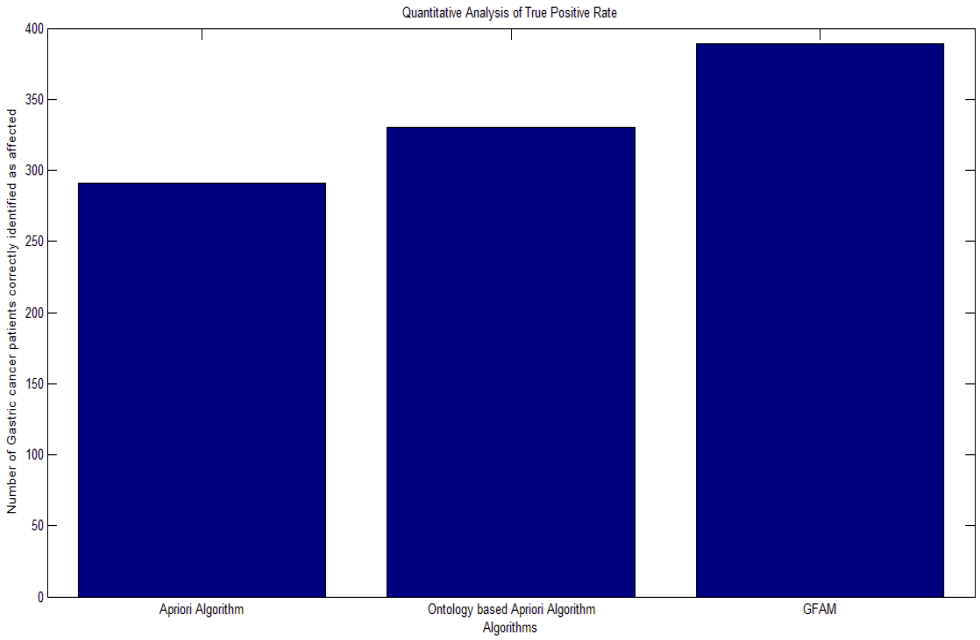

Figure-1. (a) True Positive: Gastric cancer patients correctly identified as affected.

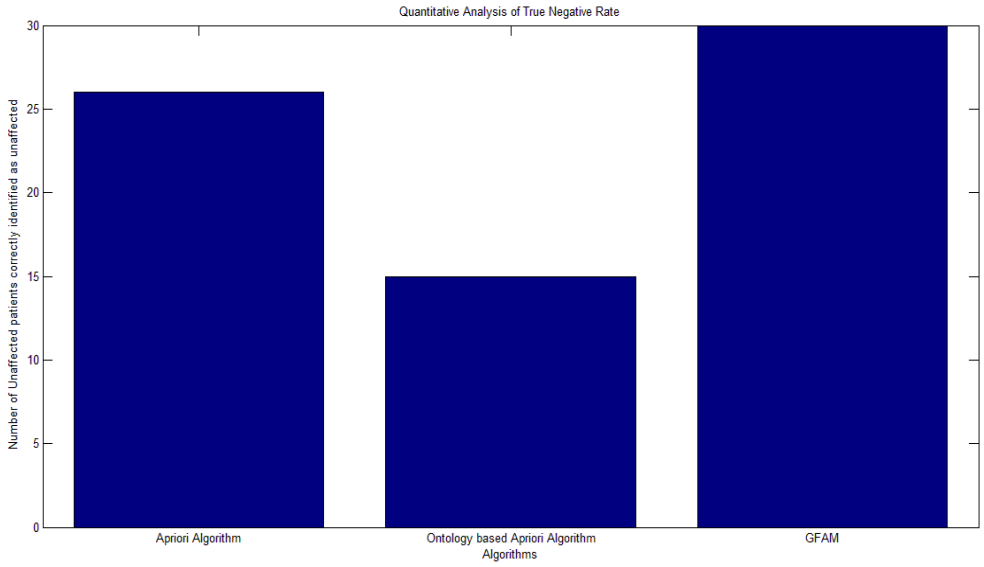

Figure-1. (b) True Negative: Unaffected patients correctly identified as unaffected.

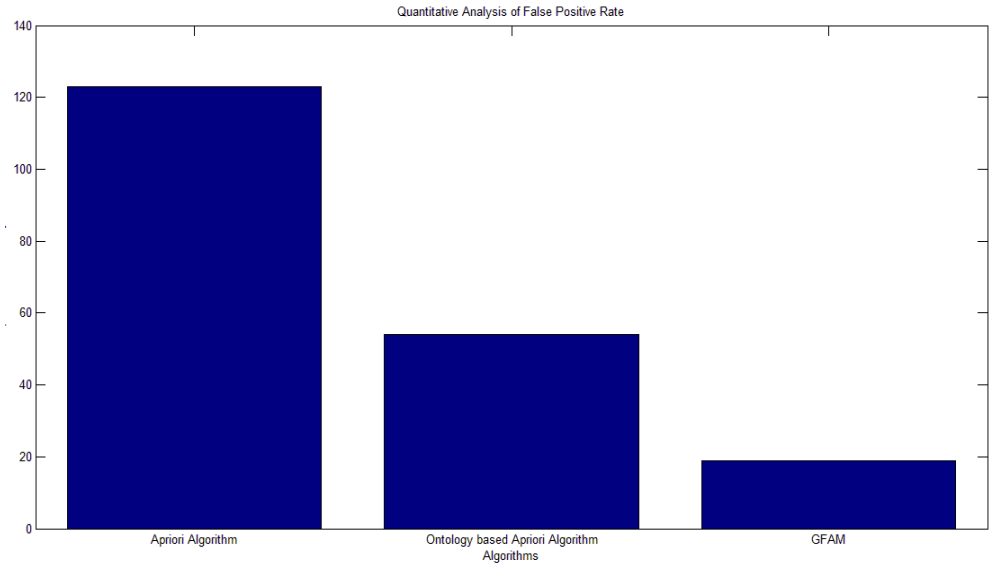

Figure-1. (c) False Positive: Unaffected patients incorrectly identified as affected.

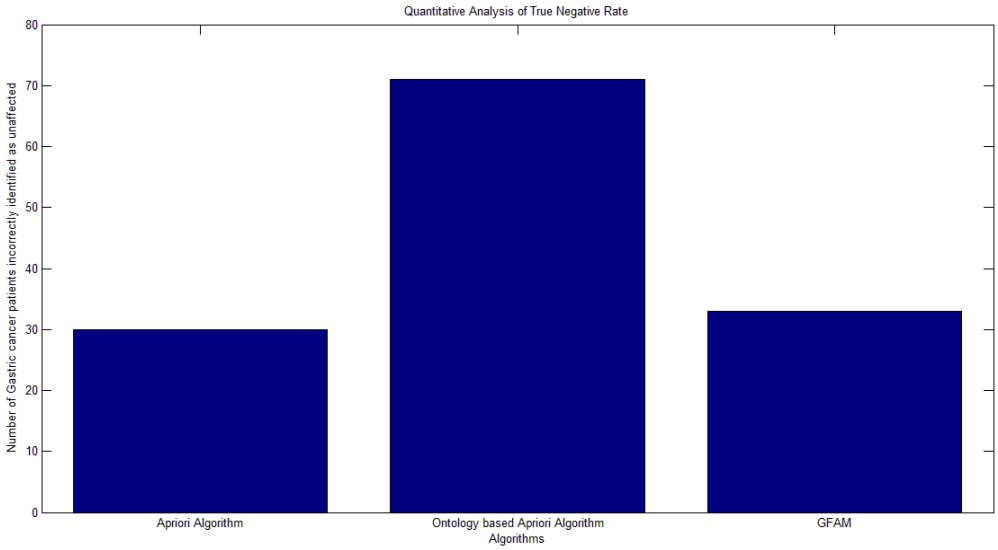

Figure-1. (d) False Negative: Gastric cancer patients incorrectly identified as unaffected.

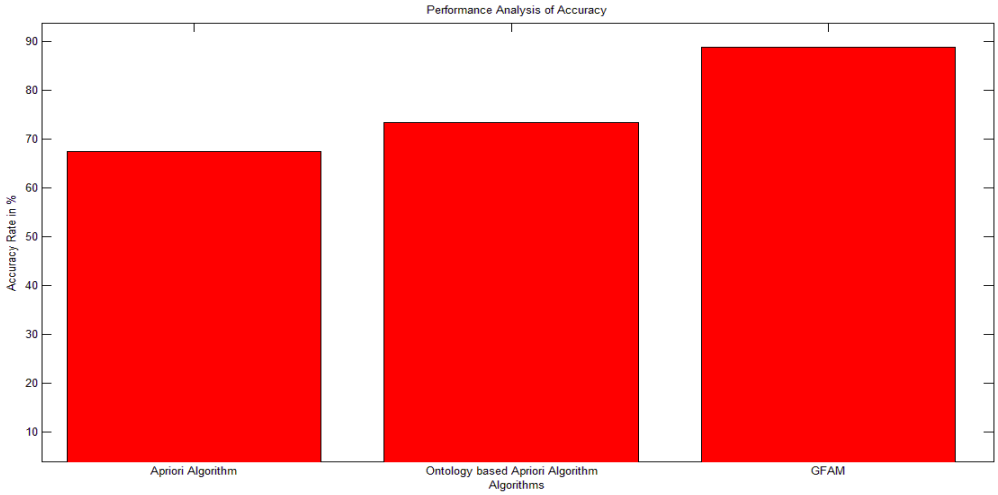

Figure-2. Performance analysis of accuracy - Apriori algorithm, Ontology based Apriori algorithm and GFAM.

It is evident that the GFAM model outperforms the other two algorithms in TP, TN and FP. But as far as FN is concerned, the GFAM has little bit performance degradation and there is scope for future work to reduce the false negative. The Figure-2 portrays the accuracy rate of the algorithms and GFAM outperforms other two algorithms and attain better classification accuracy of 88.95%. The Figure-3 depicts performance analysis of hit rate of the algorithms and it is evident that the GFAM outreaches than the two algorithms and obtains 85%. The Figure-4 exposes the performance analysis in terms of elapsed time of execution of the algorithms and it is noteworthy that GFAM consumes less time i.e. 95 seconds to classify 470 patient records. From the above performance analysis demonstrates that the GFAM model is more suitable for classifying gastric cancer patients data and there is still further scope of research in reducing the false negative rate.

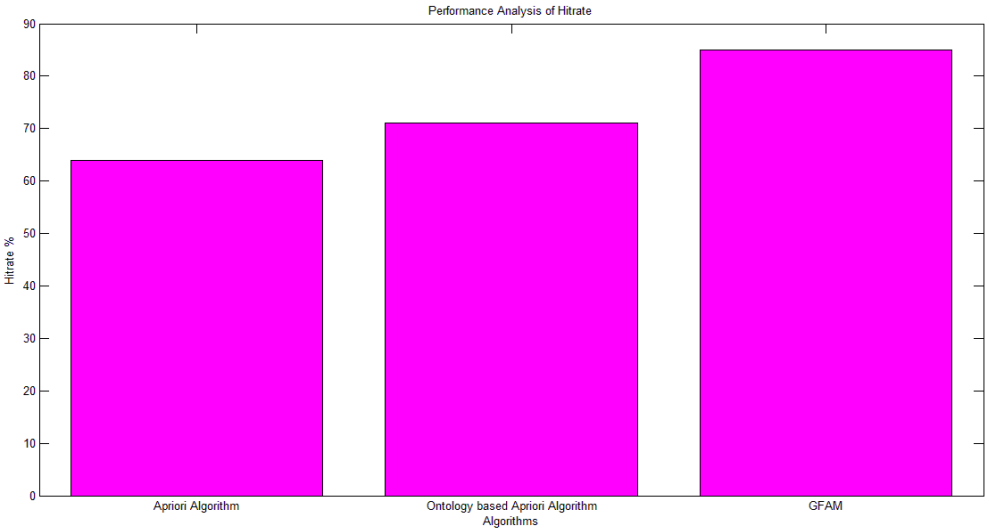

Figure-3. Performance analysis of Hit rate - Apriori algorithm, Ontology based Apriori algorithm and GFAM.

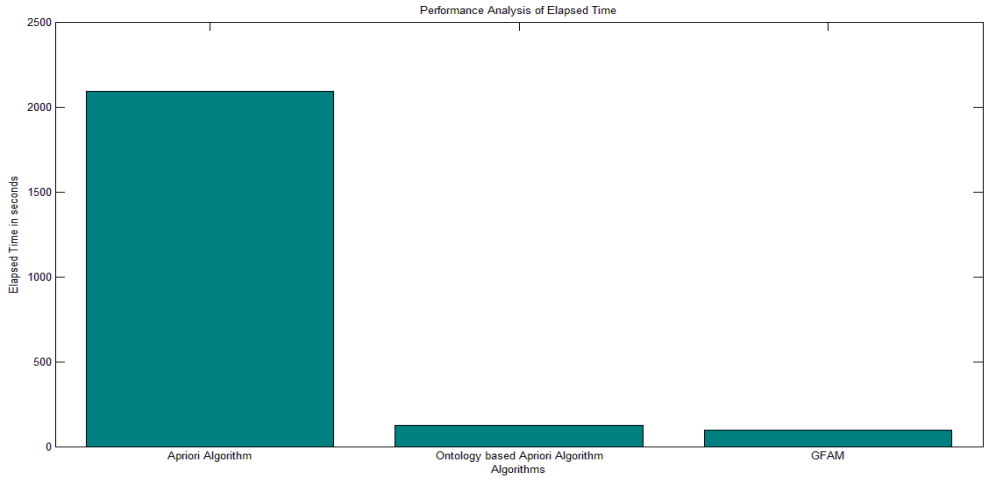

Figure-4. Performance analysis of Elapsed time - Apriori algorithm, Ontology based Apriori algorithm and GFAM.

6. CONCLUSION

In this research work, a model for gastric cancer data classification with rule extraction has been proposed. Initially, FAM and Q-learning are integrated. The resulting model is merged with a GA-based rule extractor in the next stage to fabricate GFAM. GFAM makes use of number of vital procedures. At first, GFAM generates a number of prototype nodes and allocates a Q-value to every prototype node while learning the rules. The extracted rules are capable enough to offer helpful description pertaining to the predicted class of each data sample. From the results it is evident that the GFAM classifier is capable to classify gastric cancer data in terms of accuracy and hit rate. Also the GFAM classifier consumes less time i.e. the time complexity is less than that of existing research methods.

7. FUTURE DIRECTION

Performance analysis demonstrates that the GFAM model is more suitable for classifying gastric cancer patient data and there is still further scope of research in reducing the false negative rate. As a future enhancement, it is suggested to propose fuzzy min–max neural network for data with mixed attributes. It may also be implemented with Support Vector Machine classifier.

REFERENCES

[1] Brettingham-Moore K.H., Duong C.P. 2011. Heriot A.G. using gene expression profiling to predict response and prognosis in gastrointestinal cancers-the promise and the perils. Ann SurgOncol. 18: 1484-1491.

[2] Balasubramanian S.P. 2001. Evaluation of the necessity for gastrectomy with lymph node dissection for patients with submucosal invasive gastric cancer. (Br J Surg 2001; 88: 444-9) Br J Surg. 88: 1133-1134.

[3] Boussioutas A., Li H., Liu J. 2003. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res. 63: 2569-2577.

[4] Vecchi M., Nuciforo P., Romagnoli S. 2007. Gene expression analysis of early and advanced gastric cancers. Oncogene. 26: 4284-4294.

[5] Nam S., Lee J., Goh S.H. 2012. Differential gene expression pattern in early gastric cancer by an integrative systematic approach. Int. J Oncol. 41: 1675-1682.

[6] Kim H, Eun JW, Lee H *et al.* 2011. Gene expression changes in patient-matched gastric normal mucosa, adenomas, and carcinomas. Exp Mol Pathol. 90: 201-209.

[7] Lei Z., Tan I.B., Das K. 2013. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology. 145: 554-565.

[8] Pavlidis N., Pentheroudakis G. 2012. Cancer of unknown primary site. Lancet. 379: 1428-1435.

[9] Monzon F.A., Koen T.J. 2010. Diagnosis of metastatic neoplasms: molecular approaches for identification of tissue of origin. Arch Pathol Lab Med. 134: 216-224.

[10] Greco F.A., Spigel D.R., Yardley D.A. 2010. Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction. Oncologist. 15: 500-506.

[11] Hainsworth J.D., Rubin M.S., Spigel D.R. 2013. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. J Clin Oncol. 31: 217-223.

[12] Mahmoodi SA, Mirzaie K, Mahmoudi SM. 2016. A new algorithm to extract hidden rules of gastric cancer data based on ontology. Springer Plus. 5: 312. doi: 10.1186/s40064-016-1943-9.

[13] RakeshAgrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases. VLDB; 487-499.

[14] Thara Lakshmipathy and Gunasundari Ranganathan. 2016. Significance of Data mining techniques in disease diagnosis and biomedical research - A survey. The IIOAB Journal. 7(1): 284-292.

[15] R.Gunasundari and L.Thara. 2016. Helicobacter Pylori infection and the associated stomach diseases: Comparative Data mining approaches for diagnosis and prevention measures. Proceedings of the IEEE International conference on Advances in computer applications. ISBN: 978-1-5090-3769-8; pp. 9-13.