Adaptive Feature Selection Method based on Particle Swarm Optimization for Gastric Cancer Prediction

L.Thara
Assistant Professor, Dept. of Computer Science,
PSG College of Arts & Science,
Ph.D Research Scholar, Dept. of Computer Science,
Karpagam University, Coimbatore, INDIA.
kltharavijay@gmail.com

Dr.R.Gunasundari
Associate Professor & Head, Dept. of Information Technology,
Karpagam University,
Coimbatore, INDIA.
gunasoundar04@gmail.com

***Abstract*— Medical / Clinical data mining attracted many researchers to work in that area. Bio medical engineers, computer scientists, analytic professionals are involved in such research area to develop an outcome which is nothing but a decision support system. Gastric cancer is one of the deadliest diseases that have more potential research scope. Only few research articles are published on gastric cancer prediction using computing and analytics. This research paper aims to propose an adaptive feature selection method based on particle swarm optimization for gastric cancer prediction. Performance metrics such as accuracy, hit rate and time taken for classification are taken into account for comparing the proposed AFS-PSO with the existing algorithms. 1127 real-time patients' records were obtained and the implementation of AFS-PSO is carried out using MATLAB and the results portray that AFS-PSO outperforms the existing algorithms in terms of accuracy, hit rate and elapsed time.**

Keywords*— Data mining, particle swarm optimization, machine learning, gastric cancer, real time dataset, decision support system.

I. INTRODUCTION

Data mining has improved to a wider range particularly in health care and analytics that will help medical practitioners by providing a decision support system. Decision support systems make use of health data to analyze for identifying consoles and best practices that leads to better patient care. This mechanism also cut down cost factor. There are many reviews in internet and other statistical forums stating that cancer is the deadliest disease which leads to death even in developed nations by the year 2025. For the most part, gastric cancer is taking up the fourth-most common cancer position and stands at the second leading cause of cancer deaths world-wide. This paves the motivation of this research work. Data mining algorithms and machine learning algorithms plays a significant role in clinical / medical decision support system. It is a ground truth that feature selection will help improving the accuracy of the classifier. This research work aims in design and development of adaptive feature selection method based on particle swarm optimization for gastric cancer prediction. The objective of the proposed classifier is to improve accuracy, hit rate and to reduce the time taken for classification.

II. BACK GROUND

There exist some methods that had successfully been recognized to choose, split, and categorize subtypes of gastric cancer (GC) and to figure out a few symptomatic predicaments [1]. Nam et al. [2] likewise distinguished a 973-gene signature to divide EGC from normal tissue making use of the microarray information from the coordinated tumor and neighboring non-cancerous tissues of 27 EGC patients [2]. Both unending gastritis (ChG) and intestinal metaplasia (IM) are integrated in middle of the road phase of GC, the preceding portrayed by a mitochondria-related gene expression signature at the same time as the last depicted by markers of multiplication. Since ChG has mitochondria gene expression signature, it possibly will undergo enthusiastic test whether such a signature is recognized with the metabolic subtype signature of GC [3]. Cancer of unknown primary site (CUP) is a very much perceived clinical issue, representing 3–5% of all dangerous epithelial tumors. Glass can be recognized in light of preserved tissue-particular gene expression [4]. Kirshners et al. [5] used a data set including 819 samples (24 positive and 795 negative samples) and 31 features and three algorithms CN2 Rules and C4.5 and Naive Bayes to diagnose stomach cancer. The results showed that sex and protein HER-1, are important factors in the diagnosis and classification of gastric cancer. Experimental results showed the average sensitivity >50 and 86–100 % at most, at the same time, having classification accuracy and specificity close to 65–70 %. Silvera et al. [6] used classification tree analysis to analyze data from a population-based case–control study (1095 cases, 687 controls) conducted in Connecticut, New Jersey, and Western Washington State. Frequency of reported gastro esophageal reflux disease symptoms was the most important risk stratification factor for esophageal adenocarcinoma, gastric cardia, and noncardia gastric, with dietary factors (esophageal adenocarcinoma, noncardia gastric), smoking (esophageal adenocarcinoma, gastric cardia), wine intake (gastric cardia, noncardia gastric), age (noncardia gastric), and income (noncardia gastric) appearing to modify the risk of these cancer sites. Wang et al. [7] used hierarchical clustering on 14 available clinical factors from three categories, i.e., the clinical background,

978-1-5090-5013-0/17/$31.00 ©2017 IEEE

immunohistochemistry data, and the caner's stage information. The results showed that that two clinical factors, Her-1 and gender, can clearly characterize and differentiate these three groups. In classifying and clustering somehow these methods derive patterns from the dataset. The classification algorithms that are used for ensemble result of diagnosis of stomach cancer are CART (Classification and Regression Tree), TSVM (Transductive Support Vector Mechanism).

III. Proposed Work

The particle swarm optimization (PSO) algorithm is a stochastic optimization process. It has many identical characteristics as like of the genetic algorithm (GA) and evolutionary algorithm (EA). PSO contains search method that depends on the inspiration of swarm intelligence in biological populations. PSO used to perform search operation for the global optima by updating its generations. The nature inspired phenomena of the bird's flocking or fish schooling paradigm paves way for developing PSO by J. Kennedy and R. Eberhart [8]. Every possible solution is referred as a particle, and the collection set of particles in any iteration is referred as a population. The particles flutter over a multi-dimensional search area with certain assigned speed and are updated by the earlier best performance of the particle and its neighbors. The initial population is usually begun up with the help of a random number generator for distributing the particles unvaryingly over the search area.

It is assumed that D represents the dimension of the search area, $x_{id}(t)$ represents the position of the i-th particle at d-th dimension and $v_i(t)$ is the speed of the i-th particle. The best previously visited position (up to time t) of the i-th particle is represented by x_i^{best} and the global best position of the swarm is denoted by x_g^{best}. The particle's speed and its new position are updated as follows:

$$\bar{v}_{id}(t+1)=v_{id}(t)+c_1.r_1.\left(x_i^{best}-x_{id}(t)\right)v_i(t)+c_2.r_2.\left(x_g^{best}-x_{id}(t)\right) \qquad \mathbf{(1)}$$

where c_1 and c_2 are learning factors, normally set as $c_1=c_2=2$ and r_1 and r_2 are random numbers between 0 and 1. The position of each particle is modified by adding its speed to the current position.

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1)\; i=1,2,...,N,\; d=1,2,...,D \qquad (2)$$

In this research work, the binary PSO is implemented as since the position of each individual particle is possibly represented using 0 or 1 which signifies whether a feature needs to be selected or not. The changes in particle speed can be explained as changes in the probability of finding the particle in one state. Even though PSO endow with a global search strategy to hit upon a better solution in the feature selection task, it has certain setbacks which are premature convergence and weakness in fine-tuning near local optimum points. In order to defend such setbacks, an adaptive feature selection method based on PSO, AFS-PSO was proposed.

In the AFS-PSO, the ending feature subset is obtained using seven stages. At the initial stage, the magnitude of the feature subset is obtained. In the second stage, the entire features are classified into analogous and non-analogous sets by making use of the correlation information of the features. From stage three to eight, the actual PSO mechanism is employed along with a precise local search strategy that takes the local information of the features into the search procedure. During the third stage, the predetermined numbers of particles are generated. In fourth stage, the particles are allowed to move to a new position based on their local best positions and the global best of the swarm. During the fifth stage, every particle searches in its local area considering the features correlation information. Incessantly in the sixth stage, the fitness of each particle is calculated. Then in step seventh, the local and the global best particles are replaced with those of the previous ones. Furthermore, steps fourth to seventh are repeated until the stopping criterion is satisfied, otherwise the algorithm is stopped and the best feature set is obtained.

A. Obtaining the number of features

This first stage makes use of the probabilistic random function that attempts to endow with a random number for determining the number of selected features in a bounded region. For this reason, a probabilistic formula (as per Eqn. 3) is employed to characterize the initial size of the feature subset.

$$l_{sf}=\frac{f-sf}{\sum_{i=1}^{l}(f-1)} \qquad (3)$$

where f denotes the number of the original features in a given dataset, sf is the number of the selected features, l represents the difference between f and k (i.e. l = $f-sf$) and l_{sf} is the probability value of determining sf as an initial number of features. It is clear that l_{sf} is maximized when sf is minimized. The initial number of features (i.e. sf) is obtained by making use of roulette wheel procedure that directly depends on the probability value l_{sf}. The sf is randomly selected in the range $[x,M]$ where $M=\varepsilon\,.\,f$ and x depends on a given dataset and generally is set to 3. Furthermore, ε is an adjustable parameter that controls M. If ε is set to about 1, then M is close to the number of the original features f, so the search space becomes larger, which clearly causes the high computational cost. Thus, ineffective feature subsets might be generated.

B. Grouping the features

The features are divided into analogous and non-analogous sets. The goal of the clustering in AFS-PSO is to locate relationships between features in which the most discrete features are possibly distributed into the newly generated particles. For doing this, AFS-PSO makes use of Pearson correlation coefficient for measuring correlation between different features as:

$$c_{ij} = \frac{\sum_{k=1}^{m}\left(x_i(k)-\bar{x}_i\right)\left(x_j(k)-\bar{x}_j\right)}{\sqrt{\sum_{k=1}^{m}\left(x_i(k)-\bar{x}_i\right)^2}\sqrt{\sum_{k=1}^{m}\left(x_i(k)-\bar{x}_i\right)^2}} \tag{4}$$

where c_{ij} is the correlation coefficient between two features i and j, m is the number of the samples, $x_i(k)$ and $x_j(k)$ denote the values of the feature vectors i and j for the k-th sample, respectively, and $\bar{x}_i$ and $\bar{x}_j$ represent the mean values of x_i and x_j vectors over all of the m samples, respectively. With the help of correlation coefficient between two features computes the similarity between the features which results in higher values mean that the two features have high similarity to each other. In contrast, lower values signify that the two features have low similarity.

Subsequent to computing the correlation coefficient for all possible combinations of features, the correlation value for each feature i is computed using:

$$cor_i = \frac{\sum_{j=1}^{f}\left|c_{ij}\right|}{f-1} if\ i \neq 1 \tag{5}$$

where f is the number of all features and c_{ij} denotes the Pearson correlation value between features i and j. A higher correlation value for a feature means that the feature has a high value of similarity to the other features, while a lower value means that the feature is more distinct among the others. In order to create two groups of features, AFS-PSO tries to segment the original feature set into two equivalent clusters. In order to carry out this, AFS-PSO sort out all the features in ascending order based on their corresponding correlation values. Those in the first half of the features have the lowest correlation values and they are put into the dissimilar group called D, while the rest of the features have higher correlation values and they are included in the second group called the similar group, S. The features available in the dissimilar group D are less correlated than those in the similar group S.

C. Initializing particles

In the proposed AFS-PSO, every particle is denoted by a binary vector. The span of the vector is equal to the number of the original features. When the value of a cell in the vector is set to 1, it mentions that the corresponding feature is selected and when the value is set to 0, it denotes that the corresponding feature is not selected. Alternatively, in AFS-PSO, the required number of features k is decided by the subset size determination step (i.e. step 1). After that, for each particle a speed vector is allotted using a random float number. The length of the speed vector is equal to the length of the particle vectors. Each cell of the speed vector is set to a random value in the range of [0 - 1].

D. Changing the particle positions

Each particle modifies its position according to its speed as follows (Eqn. 6):

$$v_{id}(t+1) v_{id}(t) + c_1.r_{i,1}.\left(x_i^{best} - x_{id}(t)\right) v_i(t) + c_2.r_{i,2}.\left(x_g^{best} - x_{id}(t)\right) \tag{6}$$

where c_1 and c_2 are learning factors and are generally set to $c_1 = c_2 = 2, r_{1,i}\ and\ r_{2,i}$ are random numbers in the range of [01], and $x_{id}(t)$ represents the position of the d-th dimension of the ith particle and $v_i(t)$ denotes the speed of the i-th particle. The best previously visited position (up to time t) of the i-th particle is represented by x_i^{best} and the global best position of the swarm is denoted by x_g^{best}. It should be noted that if the sum of accelerations causes the speed of that dimension to exceed $V_{\max}$, then the speed of that dimension is limited to $V_{\max}$ according to the following equation (Eqn. 7):

$$if\ v_{id}(t+1) \notin (v_{\min}, v_{\max})\ then\ v_{id}(t+1) = \max(\min(v_{\max}, v_{id}(t+1)), v_{\min}) \tag{7}$$

where $V_{\max}$ and $V_{\min}$ are user specific parameters (for this research work $v_{\max} = 3,\ v_{\min} = -3$).

The position of a particle is changed based on Eqn. 8:

$$\begin{gathered} s(v_{id}(t+1)) = \frac{1}{1+e^{-v_{id}}} \\ if\ r\ and < s(v_{id}(t+1))\ then\ x_{id}(t+1) = 1 \\ else\ x_{id}(t+1) = 0 \end{gathered} \tag{8}$$

When $(v_{id}(t+1))$ is larger than a random value, then its position value is represented by 1. Then again, if $(v_{id}(t+1))$ is smaller than a random number, then its position value is represented by 0 which means that its analogous feature is not selected for the next update.

E. Local search operations

For doing the local search in the proposed AFS-PSO method, two steps are conducted namely feature segmentation, and particle progress. In these steps for a given particle the “Add” and “Delete” operators are employed to improve the local search of a particle. In short, a particle utilizes the Add operator to select a desired number of features, and the Delete operator is employed to remove a desired number of existing features from the position of the particle. The local search is to choose discrete features with the lowest correlation. In the local search operator first of all the features selected by the particle are extracted. Particle progress plays a significant role in the local search process. For that reason, it is required to be in charge of the number of 1-bits in the newly generated particle. At this juncture, the numbers of analogous and non-analogous features are given by calculating the values of n_s and n_d, respectively. When the number of similar features in the newly generated particle is larger than $n_s, (X_s - n_s)$ similar features in Xs are taken away from the particle. In contrast, when the number of similar features in the generated particle is smaller than $n_s, (n_s - X_s)$ features in $(S - X_s)$ are added to the particle.

F. Calculating fitness

AFS-PSO makes use of support vector machine (SVM) classifier to assess a candidate feature subset solution. Prior to the assessment process all the features are given a number between -1 and 1. The normalization process changes the dominating features with greater numeric values to those with bounded numeric ranges. A linear normalization method is employed to scale the datasets as follows:

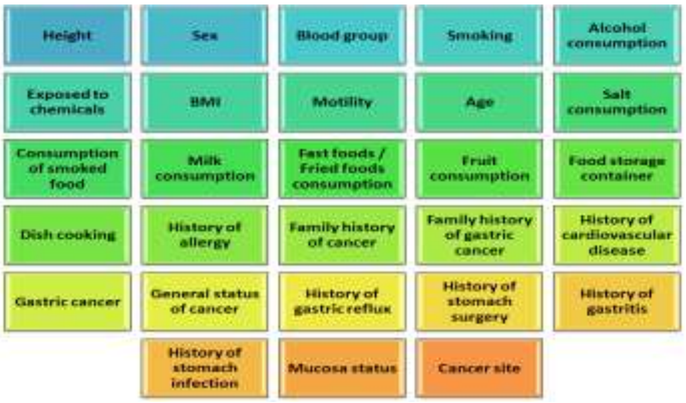
$$x^{new} = l + \left[(u - l) * \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \right) \right] \qquad (9)$$

where l and u are the lower bound and upper bound of the normalization process, respectively, $x_{\max}$, $x_{\min}$ show the maximum and minimum values of the feature x, respectively. After the normalization process, the new dataset was extracted from the (normalized) original dataset with the features that were present in the solution of the particle. Then the new dataset was divided into training and test parts (75% for training sets and 25% for the testing set). When the accuracies of two solutions were the same, then the solution using the smaller number of features were selected.

IV. Dataset and Performance Metrics

The dataset were collected from 25 healthcare centers such as hospitals and clinics in and around Coimbatore district. Around 1127 patients’ dataset were obtained with 28 features. Features are depicted in the following Figure 1.

Fig. 1. Feature set for the Patient’s Data



Height	Sex	Blood group	Smoking	Alcohol consumption
Exposed to chemicals	BMI	Motility	Age	Salt consumption
Consumption of smoked food	Milk consumption	Fast foods / Fried foods consumption	Fruit consumption	Food storage container
Dish cooking	History of allergy	Family history of cancer	Family history of gastric cancer	History of cardiovascular disease
Gastric cancer	General status of cancer	History of gastric reflux	History of stomach surgery	History of gastritis
	History of stomach infection	Mucosa status	Cancer site	

These collected features are carefully chosen as per the advice of the medical practitioners those who got potential expertise in treating gastric cancer patients. On the other hand, some of the features were not taken into account in data mining based approaches for gastric cancer diagnosis. The features include personal characteristics, behavior, systemic features and the stomach disorders / malfunctions.

The proposed AFS-PSO method is compared with two existing algorithms namely Infinite Feature Selection (IFS) [10] & Similarity Preserving Feature Selection (SPFS)[11] in terms of 3 performance metrics. Accuracy, hit rate and elapsed run time are the performance metrics taken for comparison. As far as accuracy performance evaluation is concerned, true positive(TP), true negative(TN), false positive(FP), false negative(FN) are used to compute accuracy value, as described below:

TP: Gastric cancer patients correctly identified as affected;

FP: Unaffected patients incorrectly identified as affected;

TN: Unaffected patients correctly identified as unaffected;

FN: Cancer patients incorrectly identified as unaffected.

V. Results And Discussion

True positive, true negative, false positive and false negative are quantitatively analyzed and depicted in Fig.2.

Fig. 2. (a) True Positive & (b) True Negative

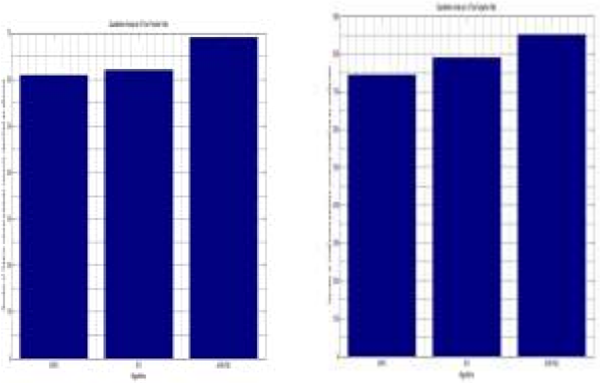
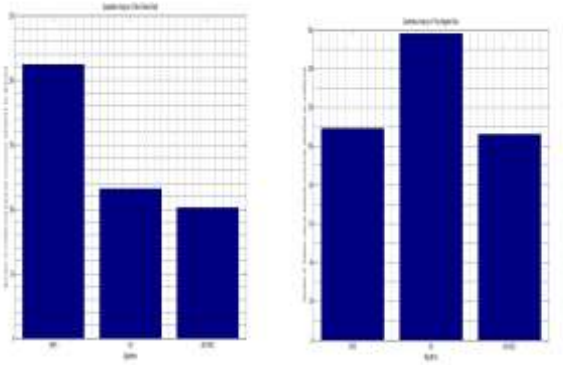


Fig. 2. (c) False Positive (d) False Negative



It is evident that the AFS-PSO model outperforms the other two algorithms in terms of TP, TN, FP and FN.

Fig.3 portrays the accuracy rate of the algorithms and AFS-PSO outperforms other two algorithms and attains better classification accuracy of 81.63%.

Fig. 3. Performance Analysis of Accuracy of the algorithms SRFS, IFS & Proposed AFS-PSO

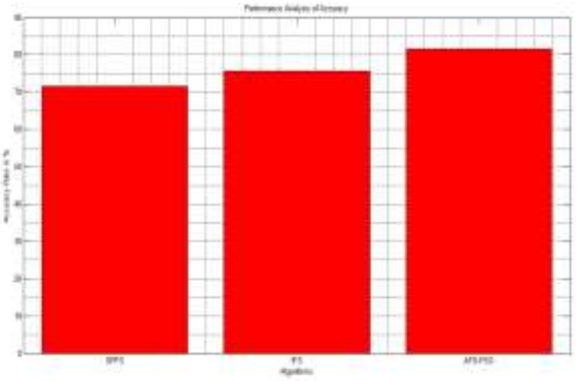
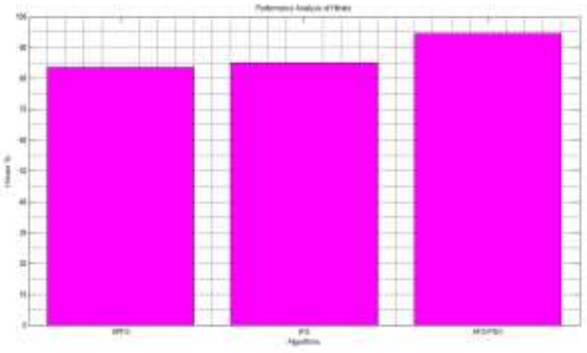


Fig.4 depicts performance analysis of hit rate of the algorithms and it is evident that the AFS-PSO outreaches than the two algorithms and obtains 94%. It exposes the performance analysis in terms of elapsed time of execution of the algorithms and it is noteworthy that AFS-PSO consumes less time i.e. 102 seconds to classify 1127 patient records.

Fig. 4. Performance Analysis of Hit rate of the algorithms SRFS, IFS & Proposed AFS-PSO



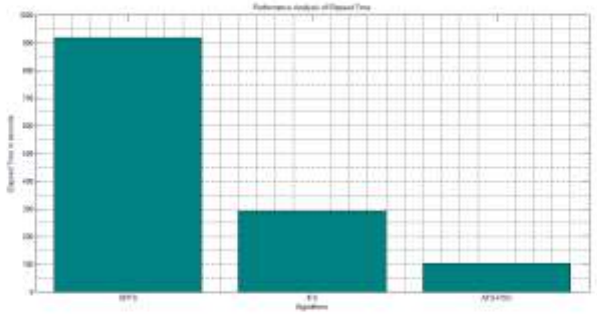
$$Hit\ Rate = \frac{correctly\ predicted\ affected\ gastric\ cancer\ patients}{Actual\ number\ of\ gastric\ cancer\ patients}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$Time\ taken = Elapsed\ time$

Fig.5 exposes the performance analysis in terms of elapsed time of execution of Algorithms.

Fig. 5. Performance Analysis of Elapsed Time of the algorithms SRFS, IFS & AFS-PSO



The above performance analysis demonstrates that the AFS-PSO attains better performance when compared with infinite feature selection mechanism (IFS) [10] and similarity preserving feature selection (SPFS) [11].

VI. Conclusion And Scope For Further Enhancement

This research work makes use of particle swarm optimization algorithm for selecting appropriate features from the dataset. Seven stages are involved in the proposed AFS-PSO which includes Obtaining the number of features, grouping the features, initializing particles, changing the particle positions, local search operations, calculating fitness. The proposed AFS-PSO is implemented in MATLAB and the results portray that AFS-PSO outperforms the other two algorithms in terms of accuracy, hit rate and elapsed time. Though the AFS-PSO model is more suitable for classifying gastric cancer patient's data, there is still further scope of research in increasing accuracy and reduce computational complexity.

References

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

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

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

[4] Pavlidis N., Pentheroudakis G.: Cancer of unknown primary site. Lancet, 379:1428–1435 (2012).

[5] Kirshners A, Parshutin S, Leja M.: Research on application of data mining methods to diagnosing gastric cancer, advances in data mining. In: Perner P (ed.) Applications and theoretical aspects. Lecture Notes in Computer Science, vol 7377. Springer-verlag, Berlin, Heidelberg, pp 24–37 (2012).

[6] Silvera SAN, Mayne ST, Marilie D, Gammon D.: Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. Ann Epidemiol 24(1) : 50–57 (2014).

[7] Wang X, Duren Z, Zhang C et al.: Clinical data analysis reveals three sub types of gastric cancer. In: IEEE 6th international conference on systems biology, pp 315–320 (2012).

[8] J. Kennedy and R. Eberhart : Particle swarm optimization. In: IEEE International Conference on Neural Networks, Perth, WA. pp. 1942-1948 vol.4 (1995).

[9] Seyed Abbas Mahmoodi, Kamal Mirzaie, Seyed Mostafa Mahmoudi.: A new algorithm to extract hidden rules of gastric cancer data based on ontology. SpringerPlus, vol.5 (2016).

[10] G. Roffo, S. Melzi, and M. Cristani.: Infinite feature selection. In Proceedings of IEEE Int. Conf. Comput. Vision, pp. 4202–4210 (2015).

[11] Z. Zhao, L. Wang, H. Liu and J. Ye.: On Similarity Preserving Feature Selection. In IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 619-632 (2013).