International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

Optimal Feature Selection Using GLFES and PSO in Imbalanced Microarray Dataset

Dr. T. Deepa

Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamilnadu, India

Abstract: ***In recent years the Biological mining has incurred the massive effort to identify the uncovered genetic variations that are associated with high-dimensional micro array data. The Imbalanced Micro array data analysis faces two complications for prediction. The first and the foremost problem is the unequal distribution of classes and the second issue is relevant feature selection and classification. The primary task in the proposed work is derived to generate the initial population by selecting the relevant features which is accomplished by Granularity learning fuzzy evolutionary sampling (GLFES). The secondary task in the proposed work is associated with optimal feature selection through PSO optimization technique. Finally the selected optimal features are classified using SVM classification.***

Keywords: Imbalanced dataset; Feature selection; GLFES; PSO

1. Introduction

Feature selection is referred as the process of selecting subset of features that are relevant for prediction analysis. It acts as a pre-processing technique which removes the redundancy and extracts the features that are more suitable for processing. It emphasis on three objectives

- Dimensionality reduction.
- Reducing the irrelevant and noisy data.
- Extracting the relevant data.

Feature selection and classification are encountered in Micro array to comprehend and discover the genetic change that causes anomalies in the normal functioning of human body. The feature selection and classification process in micro array is often influenced by its nature which is commonly large and incomplete. The dimensionality reduction and balanced class distribution is the foremost task involved in the analysis process. Thereafter, feature selection and classification is applied.

The broad spectrum of micro array analysis has stretched the inclusion of computational tools such as classification, feature selection, learning techniques to overcome the major concern for finding relevant variables, uniform class distribution and dimension reduction. Therefore, Granularity learning has high potential to handle the unequal class distribution. Moreover, to handle other issues like uncertain information processing, handling the unequal distribution and classification, fuzzy set approaches are derived. In this paper the work undergoes four methodology frames to outburst with the optimal solution.

Frame 1: Granularity learning with fuzzy→initial population for feature selection
Frame 2: Evolutionary Sampling→Balanced population
Frame 3: Particle Swarm Optimization→Optimal feature selection
Frame 4: SVM Classification→Classification

The next section discusses about the related work and proceeds with methodologies, finally with experimental result and conclusion.

2. Related Work

Particle Swarm Optimization (PSO) is a new optimization algorithm applied in numerous fields. Since, the original PSO is likely to cause the local optimization with premature convergence factor. Therefore, by means of simulated annealing algorithm, a modified algorithm is developed which makes the most optimal particle in each iteration evolving continuously, and assigns a new value to the worst particle in order to increase its disturbance (Ai-Qin Mu et al, 2009).Fast particle swarm optimization (FPSO) was proposed by (Zhihua Cui et al ,2006) to reduce numerous fitness function. The evaluation particles can change the update equations at any time they need. In the study the reliability value is evaluated using the true fitness function with threshold.

According to (Liu and Motoda, 1998; Guyon and Elisseeff, 2003; Liu and Motoda, 2007) feature selection is a process of selecting a subset of features according to certain criteria, and hence it is an important and frequently used technique in data mining for dimension reduction. Moreover, it reduces the number of features and removes irrelevant, redundant, or noisy features, and brings about obvious effects for applications by speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility.

(Juana et al., 2008) Microarray datasets, are characterized by a limited number of samples due to the cost involved in acquisition and usually a larger number of gene expressions, have highlighted the importance of the dimensionality reduction task. Too many gene expressions or features in a dataset lead to poor classification performance. Therefore, feature selection processes are essential to successfully achieve incremental classification performance on microarray datasets. Numerous methods for subset feature selection has been proposed. The benefit engaged in this method is that the possibility of the reduction of over fitting. For feature selection there exist two highly effective approaches, filters and wrappers.

Recent studies have shown that wrapper methods, which involve the performance of the underlying learning algorithm in the process of feature selection, usually result

in a selection of genes such that the classification accuracy is noticeably higher. (Xiong et al., 2001).

(Sebastian Maldonado and Richard Weber., 2009) strategy comprises wrapper method for feature selection based on sequential backward selection. The curse of dimensionality is the major problem in feature selection. The subset feature selection involves two critical concepts such as data loss that occurs due to dimensionality reduction and over fitting occurs due to sub-optimal feature selection. Hence in this study the features are removed after each iteration based on the number of errors encountered in subset validation. Then SVM classification with kernel function is applied to retrieve the optimal feature from the data set.

The related studies summarize that the Learning techniques are more suitable for feature selection and the selected feature are checked for optimality through optimization technique.

3. Methodology

The architecture of the proposed work is shown in figure 1.1.Initially, the micro array is used as input which undergoes the process of finding initial features through GLFES sampling and optimal features are selected by adopting the standard Particle Swarm optimization algorithm.

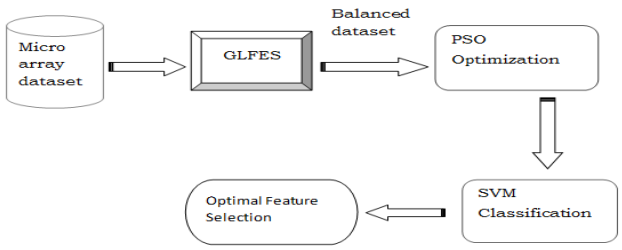

Figure 1.1: Architecture

3.1 Granule Learning and Fuzzy Evolutionary Sampling (GLFES)

Granular learning is the method that deals with appropriate gene selection. The computing is performed on the information granules rather than the data points. A granule is the process of breaking the whole process into smaller units.

The granule is the group of objects which are drawn based on the criteria. The granule can be either crisp or fuzzy. The crisp criteria involves strict cut offs which may sometimes discard the useful data. The fuzzy set focus on possibility approach which allows high-level of abstraction and deals with imprecise measurement of data. Moreover, it also concentrates on indefinite and inexact features.

The main objective of granular computing is to extract the maximum relevant and minimum redundant features from the dataset through training set labels. The fuzzy is acquired to minimize the strict cut offs that leads to removal of useful data. The granular fuzzy extracts the features based on three aspects such as High, Low and Medium.

Defnition1: In granular computing the total relevance of selected genes are expressed by

$$T_{re} = \sum_{ca} I(ca, s)$$

In the above equation were T_{re} represents the total relevance, ca is the condition attribute and s represents the sample class labels.

Definition 2: The total redundancy of the selected genes is expressed by

$$T_{rd} = \sum_{ca1,ca\,2} I(ca1, ca2)$$

In the above equation T_{rd} represents the total redundancy and ca1, ca2 are the conditional attribute of two genes in which the similarity between the two attributes are verified.

Definition 3: The maximum relevant genes are derived using the equation 6.3 in which gf represents maximum relevancy of granular computing.

$$\max(g_f) = T_{re} - T_{rd}$$

Definition 4: The fuzzy is denoted as in equation 6.4 were mf_{ij} represents the fuzzy membership function normal, malignant and diseased and A represents the attribute.

$$m(gf_{ij})^A \in [0,1].$$

The initial population is generated using granular fuzzy. The selected features are sampled using Evolutionary under sampling, Evolutionary over sampling and SMOTE Sampling.

3.2 Evolutionary Sampling (ES)

3.2.1Evolutionary under Sampling

The Evolutionary under Sampling is performed to balance the dataset by eliminating the irrelevant features of the majority class in order to equalize the majority class towards the minority class the algorithm proceeds as follows.

STEP 1: Create a random initial population P.
STEP 2: Create the chromosome C based on the number of data in the dataset
STEP 3: In the chromosome, based on the fitness $f_t = A_\rho + n_\circ / \sum_i^n = 1^{n_\circ^i}$value data's are selected and classified using SVM.
STEP 4: Continue the step 2 for total population
STEP 5: Based on the fitness function, sort the chromosome in descending order.
STEP 6: Perform the cross over using the equation 5.2 and mutation using equation 5.3 for the updated chromosome

STEP 7: Step 2 is performed and calculates the accuracy (correctly classified genes) till the end of total number of iterations.
STEP 8: Based on the final iterations result, we identify the minority and majority classes
STEP 9: Perform the under sampling, which reduces the size of the data of majority class equal to size of the data of minority class or Perform the over sampling, which replicates synthetic data to the minority class equal to size of the data of majority class or Perform the smote sampling, which adds new synthetic data to the minority class equal to size of the data of majority class.

Despite, the evolutionary sampling method generates the population that are fittest for classification. After balancing the dataset through sampling the outcome of the technique further undergoes optimality.

3.3 Particle Swarm Optimization Technique

Particle swarm optimization (PSO) is a population based stochastic optimization technique (M. Saberi et al, 2009) proposed by Dr. Eberhart and Dr. Kennedy in 1995. PSO shares many similarities with evolutionary Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optimal solution by updating generations. However, unlike Genetic algorithm, PSO does not constitute genetic operators such as crossover and mutation. In PSO, the solutions, called particles, flutter through the problem space by following the current optimum particles.

In this work, each particle of PSO keeps track of its coordinates in the problem space. The co-ordinates are the particles which are associated with the best solution called fitness. (The fitness value is also stored.) This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained by any particle related to the neighbors of the particle. This location is called lbest. Further, when a particle takes all the population of its topological neighbors, the best value is termed as global best called gbest. Formulating the best position and velocity is the key aspect in this work.

The particle swarm optimization in each iteration, changes the velocity of (accelerating) each particle toward its pbest and lbest locations. It is demonstrated that PSO gets better results in a cheaper, faster way when compared with other methods. The algorithm for PSO proceeds as follows.

Notation 1: Let $f^{\circ} = D^n \rightarrow D$ is the condition that has to be minimized fromD^n to D

Parameters: Velocity v_i and the position p_i , local bestl_b, Particle best p_b and global best g_b,(b_l, b_u) are the upper and lower boundaries of the search space and are referred as parameters of PSO.

STEP 1: For each particle i=1,2…….S initialize the particles position

$$p_i \sim \cup (b_l, b_u)$$

STEP 2: Initialize the particle best position to its initial position $p_b \leftarrow p_i$.
STEP 3: If $(f(p_b) < f(g_b))$ the swarm's best known position is updated to $g \leftarrow p_b$
STEP 4: Initialize the particle velocity

$$v_i \sim \cup (-|b_u - b_l|, |b_u - b_l|)$$

STEP 5: For each particle i= 1,2,3…….S then do
For each dimension $D_d = 1,2 \ldots \ldots \ldots \ldots n$ do
STEP 6: Extract random numbers such that $rp_b rg_b \sim \cup (0,1)$
STEP 7: Update the particle's velocity

$$v_{id} \leftarrow m(gf_{ij})^A v_{id} + gf_p r_p (p_{bd} - p_{id}) + gf_g r_g (g_{bd} - p_{id})$$

STEP 8: Update the particle's position

$$p_i \leftarrow (p_i + v_i)$$

STEP 9: if $(f(p_b) < f(g_b))$ update the swarm's best known position$g_b \leftarrow p_b$.
STEP 10: Now g_b holds the best found solution.

Thus PSO selects the optimal features for classification. It is a search heuristic method that selects the optimal feature for classification.

3.4 SVM Classification

The outcome of the balanced dataset is classified dataset using SVM classification. After classification, the data that lies on the above margin are the features that are highly satisfied. Hence these features are selected for prediction analysis. The SVM classification is used to relate the data items and to select the appropriate features for prediction.

4. Experimental Results

The experiment is carried on three datasets of micro array they are Lung cancer, Lymphoma and colon cancer. The dimensions of the dataset are 4071x96, 2026x96, 2026x96 instances and features. The performance of the proposed work is evaluated using the accuracy parameter on the sampling techniques.

Table 1.1: Result

Accuracy %			
Technique/ Dataset	Over Sampling	Under Sampling	SMOTE Sampling
Lymphoma	93.7	84.6	85.6
Lung cancer	94.3	88.6	86.6
Colon	92.4	82.5	85.9

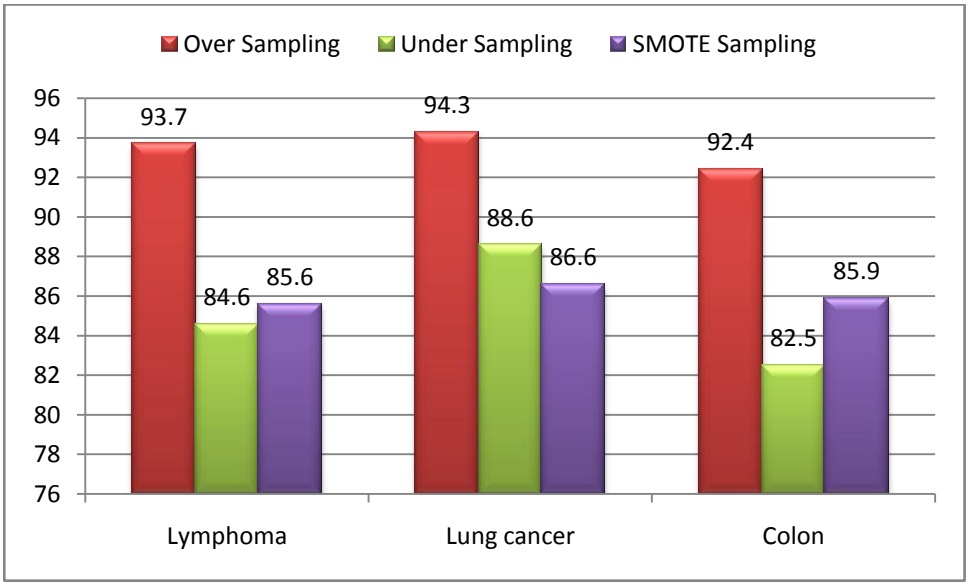

Figure 2: Accuracy

The above chart shows that the over sampling accuracy is high compared to other sampling techniques.

5. Conclusion

It is worth noting that the proposed PSO optimization attains the optimal solution that satisfies the objective of the proposed work. Comparatively, the GLFES and PSO technique gains the optimal solution through PSO. Cobbling PSO simulates the accuracy of the proposed work. Specifically, it is the standard optimization algorithm that selects the features that are both locally and globally optimal. Finally, the proposed work outshines by embedding the standard algorithms in both balancing and feature selection phases. Through GLFES the initial population is balanced and accurately derived and further PSO optimizes the accurate features.

References

[1] Alba E., J. Garcia-Nieto, L. Jourdan, and E. Talbi, 2007. Gene Selection in Cancer Classification Using PSO/SVM and GA/SVM Hybrid Algorithms, In Proc. IEEE Congress on Evolutionary Computation, 284-290.
[2] Ai-Qin Mu., De-Xin Cao and Xiao-Hua Wang, 2009. A Modified Particle Swarm Optimization Algorithm. Natural Science Research, 1(2):151-155.
[3] Bo YUAN and Wenhuang , 2012. Measure Oriented Training: A targeted approach to Imbalanced Classification problems, LIU Higher Education Press and Springer- Verlag Berlin Heidelberg, 6(5): 489-497.
[4] Brown M. P, 2000. Knowledge-based Analysis of Microarray Gene expression data by using Support Vector Machines, Proceedings of Natural Academic Science, 262–267.
[5] Blum, A.L. and P. Langley, 1997. Selection of relevant features and examples in machine learning, Artificial Intelligence, 97, 245-271.
[6] Caimiao Wei, Jiangning Li and Roger E Bumgarner, 2004.Sample size for detecting differentially expressed genes in microarray experiments, BMC Geonomics, 5:87-92.
[7] Claudio Gentile, 2003.Fast Feature Selection from Microarray Expression Data via Multiplicative Large Margin Algorithms, Advances in neural Information processing system, 1-8.
[8] Chris Ding and Hanchuan Peng, 2003. Minimum Redundancy Feature Selection from Microarray Gene Expression Data, Bio-Informatics Conference, Proceedings of IEEE, 523-528.
[9] Chuang, L.Y., H.W. Chang, C.J. Tu and C.H. Yang, 2008. Improved Binary PSO for Feature Selection Using Gene Expression Data, Computational Biology & Chemistry,32: 29-38