ISSN:0975-9646

Modified PSO Based Feature Selection for Classification of Lung CT Images

S. Sivakumar , Dr.C.Chandrasekar

Department of Computer Science, Periyar University, Salem-11, Tamilnadu, India

Abstract— **Feature selection is an optimization problem in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable recognition accuracy. Feature selection is of great importance in pattern classification, medical data processing, machine learning, and data mining applications. In this paper, continuous particle swarm optimization (PSO) is used to implement a feature selection in wrapper based method, and the k-nearest neighbor classification serve as a fitness function of PSO for the classification problem. The PSO based feature selection method is applied to the features extracted from the Lung CT scan images. Experimental results show that modified PSO feature selection method simplifies features effectively and obtains a higher classification accuracy compared to the basic PSO Feature selection method.**

Keywords— **Feature Selection, PSO, Population, Fitness function.**

INTRODUCTION

Feature selection is the problem of selecting a subset of features without reducing the accuracy of representing the original set of features. Feature selection is used in many applications to remove irrelevant and redundant features where there are high dimensional datasets. These datasets may contain a high degree of irrelevant and redundant features that may decrease the performance of the classifiers. A Feature Selection algorithm explores the search space of different feature combinations to reduce the number of features and simultaneously optimize the classification performance. In Feature Selection, the size of the search space for n features is 2^n [1] [3]. Feature selection is a multi-objective problem. It has two main objectives, which are to maximize the classification accuracy (minimize the classification error rate) and minimize the number of features. These two objectives are usually conflicting to each other and the optimal decision needs to be made in the presence of a trade-off between them.

I. PARTICLE SWARM OPTIMIZATION

PSO is an evolutionary computation technique proposed by Kennedy and Eberhart in 1995 [2] [4]. In PSO, a population, called a swarm, of candidate solutions are encoded as particles in the search space. PSO starts with the random initialization of a population of particles. The whole swarm move in the search space to search for the best solution by updating the position of each particle based on the experience of its own and its neighbouring particles [2] [3]. During movement, the current position of particle I is represented by a vector $xi = (xi1, xi2, ..., xiD)$, where D is the dimensionality of the search space. The velocity of particle i is represented as $vi = (vi1, vi2, ..., viD)$, which is limited by a predefined maximum velocity, v_{max} and $v^t_{id} \in [-v_{max}, v_{max}]$. The best previous position of a particle is recorded as the personal best pbest and the best position obtained by the population thus far is called gbest. Based on pbest and gbest, PSO searches for the optimal solution by updating the velocity and the position of each particle according to the following equations:

$$\boldsymbol{x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1}} \quad (1)$$

$$\boldsymbol{v_{id}^{t+1} = w * v_{id}^{t} + c1 * r1i * (p_{id} - x_{id}^{t}) + c2 * r2i * (p_{gd} - x_{id}^{t})} \quad (2)$$

where t denotes the t^{th} iteration, d denotes the d^{th} dimension in the search space D, w is inertia weight. c1 and c2 are acceleration constants. $r1i$ and $r2i$ are random values uniformly distributed in [0, 1]. pid and pgd represent the elements of $pbest$ and $gbest$ in the d^{th} dimension[5].

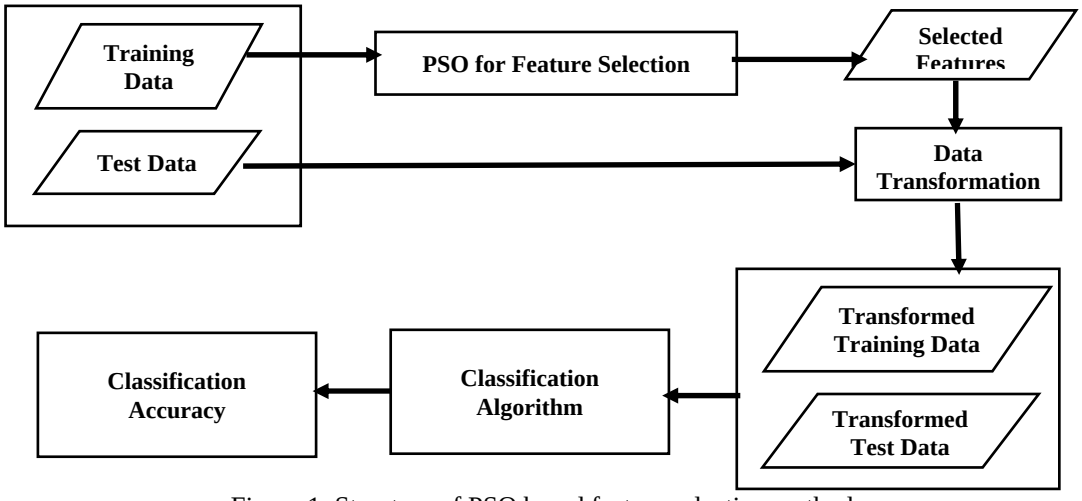

Figure 1: Structure of PSO based feature selection method

II. FEATURE SELECTION USING PSO

From the figure1, the algorithm firstly runs on the training set of the dataset to select a subset of relevant features, which is the evolutionary training process. Then the training set and the test set are transformed to a new training set and a new test set by removing the features that are not selected. A classification algorithm is trained (learns) on the transformed training set. The learnt classifier is then applied to the transformed test set to obtain the final testing classification performance [6].

A. Particle Representation:

In PSO for feature selection, the representation of a particle is a n-bit string, where n is the total number of features in the dataset. The position value in the d^{th} dimension (i.e. x_{id}) is in [0,1], which shows the probability of the d^{th} feature being selected. A threshold θ is used to determine whether a feature is selected or not. If $x_{id} > \theta$, the d^{th} feature is selected. Otherwise, the d^{th} feature is not selected [9].

B. Training Process

The training process of a PSO based wrapper feature selection algorithm is shown in Figure 2. The key step is the goodness/fitness evaluation procedure. The position of a particle represents a selected feature subset. By removing the features that are not selected, the training set is transformed to a new training set. The classification performance of the selected features is evaluated on the transformed training set. Based on the classification performance, the fitness of the particle is then calculated according to the predefined fitness function[10]. After evaluating the fitness of all particles, the algorithm updates the *pbest* and *gbest,* and then updates the velocity and position of each particle. The algorithm stops when a predefined stopping criteria, that is the maximum number of iterations or an optimal fitness value, has been met. During the training process, Equation 3, which aims to minimize the classification error rate, is used as the fitness function to evaluate the goodness of particle *i*, where the position *xi* represents a feature subset [6].

$$fitness(x_i) = Error\ Rate \quad (3)$$

Where the Error rate is determined by

$$Error\ Rate = (FP + FN)/(FP + FN + TP + TN) \quad (4)$$

Where TP, TN, FP and FN stand for true positives, true negatives, false positives and false negatives, respectively.

The adaptive functional values were data based on the particle features representing the feature dimension; this data was classified by a k-Nearest Neighbor (kNN) to obtain classification accuracy; the kNN serves as an evaluator of the PSO fitness function. For example, when a 8-dimensional data set (n=8) $S_n = F_1F_2F_3F_4F_5F_6F_7F_8$ is analyzed using particle swarm optimization to select features smaller than n.

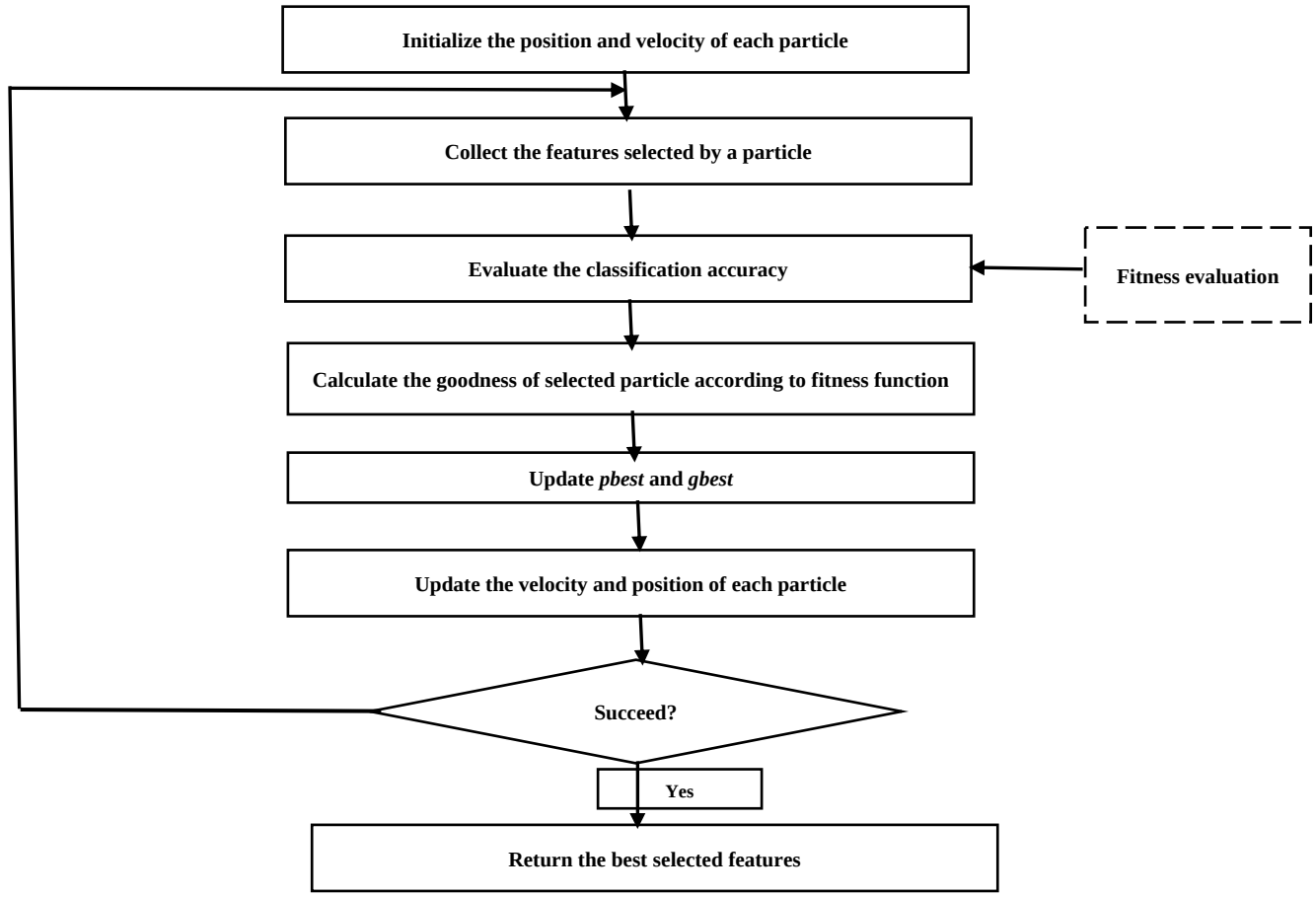

Figure 2: Training process of a PSO based wrapper feature selection algorithm

The following pseudo code shows the basic PSO Feature Selection process [5] [6].

Basic PSO algorithm for Feature Selection

Input : Training Data set and a Test Data set;

Output : Selected feature subset

```
1 Begin
2     randomly initialize the position and velocity of each particle;
3     while Maximumiterations is not reached do
4         evaluate fitness of each particle ; /* according to (3) */
5         for i=1 to PopulationSize do
6             update the pbest of particle i;
7             update the gbest of particle i;
8         for i=1 to PopulationSize do
9             for d=1 to Dimensionality do
10                update the velocity of particle i according to (2);
11                update the position of particle i according to (1);
12 calculate the classification accuracy of the selected feature subset on the test set;
13 return the position of gbest (the selected feature subset);
```

III. MODIFIED PSO FEATURE SELECTION

In the Modified PSO Feature Selection, not only focusing on fitness value, and also the number of features also considered. The following two criterions also checked in the modified PSO.

(1) If *Fitness (xi) = Fitness (pbest) and* |*xi*| < |*pbest*| **then**
pbest = xi ; // Update the pbest of particle i

(2) If any *Fitness(pbest) = Fitness(gbest) and* |*pbest*| < |*gbest*| **then**
gbest = pbest ; // Update the gbest of particle i

The above two conditions are used to select the optimum features with highest classification accuracy. The following pseudo code shows the modified PSO Feature Selection.

Modified PSO Feature Selection Algorithm

Input : Training data set and Test data set

Output : Selected feature subset

```
1 Begin
2     randomly initialize the particles position and velocity;
3     while Maximum Iterations is not reached do
4         evaluate the fitness (classification performance) of each particle on the Training set;
5         for i=1 to Population Size do
              // Fitness(xi) measures the error rate of xi
6             if Fitness(xi) < Fitness(pbest) then
7                 pbest = xi ; // Update the pbest of particle i
8             else if Fitness(xi) = Fitness(pbest) and |xi| < |pbest| then
9                 pbest = xi ; // Update the pbest of particle i
10            if any Fitness(pbest) < Fitness(gbest) then
11                gbest = pbest ; // Update the gbest of particle i
12            else if any Fitness(pbest) = Fitness(gbest) and |pbest| < |gbest| then
13                gbest = pbest ; // Update the gbest of particle i
14        for i=1 to Population Size do
15            update the velocity and the position of particle i;
16 calculate the classification accuracy of the selected feature subset on the Test set;
17 return the position of gbest (the selected feature subset);
```

TABLE I
PERFORMANCE ANALYSIS OF PSO AND MODIFIED PSO FEATURE SELECTION

	Unreduced Data		PSO based Feature Selection		Modified PSO Feature Selection	
	# of Features	Accuracy (%)	# of Features	Accuracy (%)	# of Features	Accuracy (%)
First Order Statistical Features	5	73.56	4	80.21	4	81.05
GLCM based Features	14	76.30	11	82.76	9	88.49
GLRLM based Features	7	69.83	5	79.43	4	87.64

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of the modified PSO for Feature Selection the LIDC-IDRI Lung CT scan images were used. The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions. It is a web-accessible international resource for development, training, and evaluation of computer-assisted diagnostic (CAD) methods for lung cancer detection and diagnosis. Each study in the dataset consist of collection of slices and each slice of the size of 512 X 512 in DICOM format. The lungs image data, nodule size list and annotated XML file documentations can be downloaded from the National Cancer Institute website [7][8]. For the experiment we taken 126 Non-Cancer Lung CT scan images and 280 Cancer Lung CT images from the LIDC dataset.

All the CT scan images are preprocessed through wiener filter and the lung portion is extracted through morphological operations. From the segmented lung portion, both the first order statistical features (mean, variance, standard deviation, skewness, and kurtosis) and second order statistical features (GLCM based 14 Haralick features and GLRLM based 7 features) are extracted. These features are taken as the input for both the PSO and modified PSO based Feature Selection. In order to evaluate the fitness function (Error Rate) we uses kNN classifier with k=1 and the PSO parameters are set to Maximum-iterations=300, Population Size=100, c1=2, c2=2, interia_weight w= (maximum_iteration-current_iteration_count)/maximum_iteration, and θ=0.5 with the average run of 6 times.

Table 1 shows the three different type of features which are extracted from the Lung CT scan images namely first order statistical features, GLCM based Haralick features and GLRLM based features used in the experiment. The three features sets have different number of features (5, 14, 7), with two classes and instances as the representative samples of the problems that the proposed algorithms can address. In the experiments, the instances in each dataset are randomly divided into two sets: 65% as the training set and 35% as the test set. From table1, the modified PSO Feature selection yields better accuracy with minimal set of features.

V. CONCLUSIONS

Building an efficient classification model for classification problems with different dimensionality and different sample size is important. The main tasks are the selection of the features and the selection of the classification method. In this paper, we used modified PSO feature selection to perform feature selection and then evaluated fitness values with a kNN. Experimental results show that our method simplified feature selection and the total number of parameters needed effectively, thereby obtaining a higher classification accuracy compared to the basic PSO feature selection method. The proposed method can serve as an ideal pre-processing tool to help optimize the feature selection process, since it increases the classification accuracy and, at the same time, keeps computational resources needed to a minimum.

ACKNOWLEDGMENT

The First Author extends his gratitude to UGC as this research work was supported by Basic Scientist Research (BSR) Non-SAP Scheme, under grant reference number, F-41/2006(BSR)/11-142/2010(BSR) UGC XI Plan.

REFERENCES

[1] K. Waqas, R. Baig, and S. Ali, "Feature subset selection using multiobjective genetic algorithms," in IEEE 13th International Conference on Multitopic Conference (INMIC'09), pp. 1–6, 2009.

[2] L. Ke, Z. Feng, Z. Xu, K. Shang, and Y.Wang, "A multiobjective ACO algorithm for rough feature selection," in Second Pacific-Asia Conference on Circuits, Communications and System (PACCS), vol. 1, pp. 207–210, 2010.

[3] P. Engelbrecht, Computational intelligence: an introduction (2. ed.). Wiley, 2007.

[4] S. Yang, L. Y. Chuang, C. H. Ke, and C. H. Yang, "Boolean binary particle swarm optimization for feature selection," in IEEE Congress on Evolutionary Computation (CEC'08), pp. 2093–2098, 2008.

[5] Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," European Journal of Operational Research, vol. 206, no. 3, pp. 528–539, 2010.

[6] Chakraborty, "Feature subset selection by particle swarm optimization with fuzzy fitness function," in 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE'08), vol. 1, pp. 1038–1042, 2008.

[7] S.Sivakumar and C.Chandrasekar, "Lung Nodule Segmentation through Unsupervised Clustering Models", Procedia Engineering, vol.38,p. 3064-3073.

[8] S.Sivakumar and C.Chandrasekar, "Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines", International Journal of Engineering and Technology, vol. 5, no. 1, pp. 179-185, 2013.

[9] G. Talbi, L. Jourdan, J. Garcia-Nieto, and E. Alba, "Comparison of population based metaheuristics for feature selection: Application to microarray data classification," in IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'08), pp. 45–52, 2008.

[10] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143–159, 2002.