International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.6, pp : 781-784 1 June 2014

Feature Selection Using ABC forthe Lung CT Scan Images

S. Sivakumar, Dr.C.Chandrasekar

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

ssivakkumarr@yahoo.com, ccsekar@gmail.com

Abstract:Feature Selection is an important preprocessing step for most machine learning algorithms especially pattern classification. Feature Selection aims in determining the most relevant and useful subset of features from the dataset representing any application domain, without compromising the predictive accuracy represented by the actual set of features. There are many meta-heuristic search algorithms are usedto solving combinatorial optimization problems. This paper aims at investigating, implementing, and analyzing a feature selection method using the Artificial Bee Colony approach to classification of lung cancer image database.

Keywords: feature selection, ABC, k-NN, SVM, classification accuracy.

Introduction:

Feature selection is an important step used in several tasks, such as image classification, cluster analysis, data mining, pattern recognition, image retrieval, among others. It is a crucial preprocessing technique for effective data analysis, where only a subset from the original data features is chosen to eliminate noisy, irrelevant or redundant features. This task allows to reduce computational cost and improve accuracy of the data analysis process [1] [2] [3]. There are two major categories of feature selection process namely filter and wrapper models. Many evolutionary algorithms have been used for feature selection, which include genetic algorithms and swarm algorithms [4]. Swarm algorithms include, in turn, Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Bat Algorithm (BAT), and Artificial Bee Colony [7] [8].

The artificial bee colony (ABC) algorithm is a new population-based metaheuristic technique based on the foraging behavior of honey bee swarm. The ABC algorithm was initially developed by Karaboga in 2005. On the basis of their foraging behavior, real bees are divided into three categories—employed, scouts and onlookers. Employed bees are those bees that are currently exploiting the food sources [5]. All employed bees are responsible for bringing loads of nectar from their food sources to the hive and sharing information about their food sources with onlookers. Onlookers are those bees that are waiting in the hive for employed bees to share information about their food sources. The employed bees share the information about their food sources with onlookers by dancing in a common area. The nature and duration of the dance of an employed bee depends on the quality of the food source currently being exploited by it. Onlookers watch numerous dances of employed bees before selecting a food source. The probability of selecting a food source is directly proportional to its quality. Therefore, elite food sources attract more onlookers than the poor ones. Scouts are those bees which are exploring a new food source in the vicinity of the hive. Whenever a scout bee or an onlooker bee finds a food source it becomes employed. Whenever a food source is fully exploited, the associated employed bee abandons it and becomes a scout. As soon as this scout discovers a new food source in the vicinity of its hive, it again becomes employed. Hence, employed and onlooker bees do the job of exploitation, whereas scouts do the job of exploration.

In ABC algorithm, the artificial beecolony consists of three categories of bees namely employee, onlooker and scout bees. A bee waiting on the dance area for making decision to choose a food source is called an onlooker and a bee going to the food source visited by itself previously is named an employed bee. A bee carry out the random search is called the scout. In the ABC algorithm, first half of the colony consists of employed artificial bees and the second half constitutes the onlookers. The count of employed bees are equal to the number of food sources (SN). The employed bee whose food source has been exhausted becomes a scout bee [6] [9].

The search carried out by the artificial bees can be summarized as follows:

- Each employed bee governs a food source within the neighborhood of the food source in her memory and estimates its profitability.
- Each employed bee shares information with onlookers waiting in the hive and then each onlooker selects a food source site depending on the information given.
- Each onlooker determines a food source within the selected site chosen by herself and evaluates its profitability.
- An employed bee of which the source has been abandoned becomes a scout and starts to search a new food source randomly.

In the ABC algorithm, the possible solutions of the optimization problem are represented as the position of the food sources and the fitness of the associated solution is corresponded to the nectar amount of the food source. The number of the employed bees is equal to the number of food sources being exploited at the moment or to the number of solutions in the population.

Major steps involved in the ABC algorithm:

1. Initialize the population of solutions x

2. Evaluate the population

3. Iteration = 1

4. Repeat

5. Produce new solutions (food source positions) $V_{i,j}$ in the neighborhood of $X_{i,j}$for the employeed bees using the formula $V_{i,j}=X_{i,j} + \Phi_{i,j} (X_{i,j}-X_{k,j})$ (where k is the random solution of neighbor of I, Φ is a random number between -1 to +1) and evaluate them.

6. Apply the greedy selection process between X_i and V_i

7. Calculate the probability values P_i for the solutions X_i by means of their fitness valuesusing the following equation:

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i}$$

In order to calculate the fitness values of solutions using the following equation:

$$fit_i = \left\{ \frac{1}{1+f_i} \qquad if\ f_i \geq 0 \right\} \text{(or)}$$

$$fit_i = \{1 + abs(f_i) \quad if\ f_i < 0\}$$

8. Produce the new solutions (new positions) V_i for the onlookers from the solutions X_i selected depending on P_i and evaluate them

9. Apply the greedy selection process for the onlookers between X_i and V_i

10. Determine the abandoned solution (source), if exists, and replace it with a new randomly produced solution X_i for the scout using the following equation:

$X_{i,j} = min_j + rand(0,1)*(max_j-min_j)$

11. Memorize the best food source position (solution) achieved so far
12. Iteration = Iteration + 1
13. Until Iteration = Maximum Iteration Number (MIN)

The quality of the solution represented bythat food source is corresponds to the nectar amount of the food source. Onlookers are placed onto the food sources by using "roulette wheel selection" method. Every bee colony has scouts that are the colony's explores. The explorers do not have any guidance while looking for food. They are primarily concernedwith finding any kind of food source. As a result of such behavior, the scouts are characterized by low search costs and a low average in food source quality. Occasionally, the scouts may accidentally discover rich entirely unknown food sources. In the case of artificial bees, the artificial scouts might have the fast discovery of the group of feasible solutions as a task. In ABC algorithm, one of the employed bees whose food source has been exhausted is selected and classified as the scout bee. The classification is controlled by a control parameter called limit. If a solution representing a food source is not enriched until a fixed number of trials, then that food source is abandoned by its employed bee and the employed bee becomes a scout. The number of trials for releasing a food source is equal to the value of limit, which is an important control parameter of ABC algorithm.

Feature selection using ABC:

The possible solutions to the problem can be represented by vectors with real values, the candidate solutions to the feature selection problem are represented by bit vectors. Each food source is associated with a bit vector of size N, where N is the total number of features. The number of features are represented as a vector and the position of the vector is to be evaluated. If the value at the corresponding position is 1, this indicates that the feature is part of the subset to be evaluated. On the other hand, if the value is 0, it indicates that the feature is not part of the subset to be assessed. Additionally, each food source stores its quality (fitness), which is given by the accuracy of the classifier using the feature subset indicated by the bit vector.

The main steps of the ABC based feature selection method are described as follows:

Step 1: Create initial food sources:
For feature selection, it is desirable to search for the best accuracy using the lowest possible number of features. For this reason, the proposed method follows the forward search strategy. The algorithm is initialized with N food sources, where N is the total number of features. Each food source is initialized with a bit vector of size N, where only one feature will be presented in the feature subset, that is, only one position of the vector will be filled with 1.

Step 2: Calculate the accuracy of the selected subset using classifier and use accuracy as fitness: The feature subset of each food source is submitted to the classifier, and accuracy is stored as the fitness of food source.

Step 3: Regulate neighbors of chosen food sources by employed bees using modification rate (MR)parameter:
Each employed bee visits a food source and explores its neighborhood. For feature selection, a neighbor is created from the bit vector of the original food source. For each position of the bit vector or feature, a random and uniform number R_i is generated in the range between 0 and 1. If this value is lower than the MR, the feature is inserted into the subset, that is, the vector value at that position is filled with 1. Otherwise, the value of the b it vector is not modified.

Step 4: Submit a feature subset of neighbors to the classifier and use accuracy as fitness:
The feature subset created for each neighbor is submitted to the classifier, and accuracy is stored as the neighbor's fitness.

Step 5: Evaluate the fitness of neighbors:
If the food source quality of the newly created neighbor is better than the food source under exploration, then the neighbor food source is considered as a new one and information about its quality will be shared with other bees. Otherwise, variable LIMIT, from the food source where the neighborhood is being explored, is incremented. If the value of MAX_LIMIT is lesser

than the LIMIT, then the food source is abandoned, that is, the food source is exhausted. In other words, the employed bees explored a food source neighborhood MAX LIMIT times; however, they did not find any food source with better quality, such that it is not worthwhile following a way where all food sources around it have worse quality than the current source. For each abandoned source, the method creates a scout bee to randomly search a new food source.

Step 6: Distribution of onlookers: onlooker bees collect information about the fitness of food sources visited by employed bees and choose food sources with either better probability of exploration or better fitness. At the moment that onlooker bees choose the food source to be explored, they become employed bees and execute step 3.

Step 7: Memorize the best food source:
After all onlookers have been distributed, the food source with the best fitness is stored.

Step 8: Find abandoned food sources and produce new scout bees:
For each abandoned food source, a scout bee is created and a new food source is generated, where a bit vector with size N of features is randomly created and submitted to the classifier, and accuracy is stored. The new food source is assigned to scout bees, and then they become employed bees and execute step 3.

k-Nearest-Neighbor classifier:

k-Nearest-Neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. k-Nearest-Neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. It uses Euclidean distance as the distance measure. For evaluate the accuracy of the selected subset of features, the kNN classifier acts as the fitness evaluator [7].

Support Vector Machines:

SVM is a machine learning tool, based on the idea of data classification. It performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. The separation of data can be either linear or non-linear. Kernel function maps the training data into a kernel space and the default kernel function is the dot product. For non-linear cases, SVM uses a kernel function which maps the given data into a different space; the separations can be made even with very complex boundaries. The different types of kernel function include polynomial, RBF, quadratic, Multi-Layer Perceptron (MLP). Each kernel is formulated by its own parameters like γ, σ, etc. By varying the parameters the performance rate of the SVM can be measured.

Numerous recent studies have stated that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification performance than the other data classification algorithms.SVMs are set of associated supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes.

Training vectors x_i are mapped into a complex (may be infinite) dimensional space by the function Φ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space .C > 0 is the penality parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \Phi(x_i)^T . \Phi(x_j)$ is called the kernel function. There are several kernel functions in SVM, so how to select a good kernel function is also a research issue. However, for general purposes, there are some popular kernel functions [10]:

(1) Linear kernel: $K(xi , xj) = x_i^T x_j$

(2) Polynomial kernel: $K(x_i , x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

(3) RBF kernel : $K(x_i , x_j) = \exp(-\gamma \| x_i - x_j \|^2) , \gamma > 0$

(4) Sigmoid kernel: $K(x_i , x_j) = \tanh(\gamma x_i^T x_j + r)$

Here, γ, r and d are kernel parameters. In the above kernel function types, 'γ' and 'd' are the kernel parameters, whose values are 1 and 3 respectively.

Results and Discussion:

In order to evaluate the performance of the ABC based 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 150 Non-Cancer Lung CT scan images and 250 Cancer Lung CT images from the LIDC dataset.

TABLE 1: PARAMETERS USED FOR THE ABC BASED FEATURE SELECTION

Parameters	Value
Colony Size(NP) (Employee bees+ Onlooker bees)	200
Food Sources(FS)	100
Cycles	100
Parameters of the problem (D)	5, 13, 7, 8, 7

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 GLDM based 8features and GLGLM based 7 features) are extracted. These features are taken as the input for ABC based Feature Selection. In order to evaluate the fitness of both the employee phase and onlooker phase we uses kNN classifier with k=1. The table1, shows the parameters used for the algorithm and the table2, shows the selected features in different runs.

Table 2: List of Selected Features by ABC with K-NN

Run	First Order Features	GLCM	GLRLM	GLDM	GLGLM
1	[1,4]	[1,2,3,8,9,10,11]	[1,2,7]	[1,3,6]	[2,5,7]
2	[2,5]	[1,2,4,5,6,9,10]	[1,2,4,6]	[1,3,6,7]	[3,4,5]
3	[2,4,5]	[1,2,5,7,11,13]	[1,2,5,7]	[2,3,6,8]	[1,3,5,7]
4	[1,3]	[1,2,4,6,8,10,13]	[1,2,3,5]	[1,4,5,7]	[1,3,5,6]
5	[1,2,3]	[1,3,4,7,9,11,12]	[1,2,5,7]	[1,2,3,7]	[2,3,4,7]
6	[1,2]	[1,3,7,10,12,13]	[1,3,5]	[1,4,6,7]	[2,3,5,6,7]
7	[1,4]	[1,4,5,7,10,12]	[1,5,7]	[1,2,8]	[2,3,5,6,7]
8	[1,5]	[1,2,8,10,13]	[1,3,5,7]	[1,4,5,6]	[1,2,5,7]
9	[1]	[1,5,8,9,10,12]	[1,2,5,6]	[3,4,5,7,8]	[3,4,6,7]
10	[1,2,4]	[1,3,5,7,8,11,13]	[1,2,6]	[3,4,5,6]	[1,3,4,6]

For the experiments, the instances in each dataset are randomly divided into two sets: 70% as the training set and 30% as the test set.

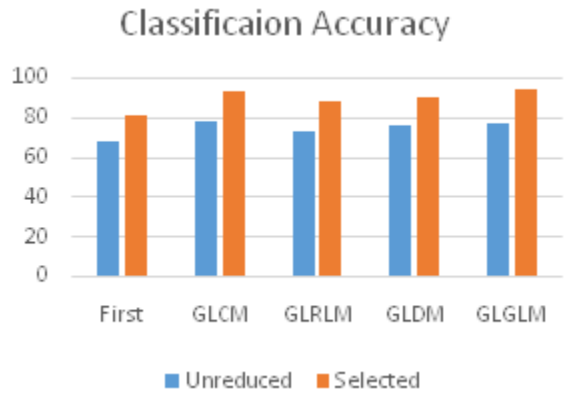

Figure 1: Classification accuracy using SVM

The figure 1 shows the ABC based Feature selection with k-NN as the fitness evaluator yields better accuracy with minimal set of features in GLGLM based features and GLCM based features.

Conclusion:

In this paper, ABC based feature selection for the features extracted from the Lung CT scan images, in order to classifying the image into cancerous or non-cancerous. The ABC based feature selection minimize the feature subset with relevant features. In order to evaluate the accuracy of the selected features, the linear kernel based support vector machine algorithm is used as the classifier for the selected subset. The ABC with k-NN selects the minimal number of features with the highest classification accuracy where compare with the unreduced feature set.

Acknowledgement:

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

References:

i. *Ahmed, E.F., W.J. Yang and M.Y. Abdullah, "Novel method of the combination of forecasts based on rough sets", J. Comput. Sci., 5: 440-444.2009.*

ii. *ShiXin Yu, Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data, Ph.D. Thesis, The University of Antwerp, 2003.*

iii. *Dash.M and Liu.H., "Feature Selection for Classification", Intelligent Data Analysis, Vol.39, No. 1, pp. 131 – 156, 1997.*

iv. *Mohammed El-Abd, "A Cooperative Approach to the Artificial Bee Colony Algorithm", IEEE, 2010.*

v. *D. Karaboga, B. Basturk, "Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems", LNCS: Advances in Soft Computing: Foundations of Fuzzy Logic and Soft Computing, Springer – Verlag, IFSA 2007, vol. 4529/2007, pp.789-798.*

vi. *WenpingZou, Yunlong Zhu, Hanning Chen and Zhu Zhu, Cooperative Approaches to Artificial Bee Colony Algorithm, Proc. IEEE International Conference on Computer Application and System Modeling, Vol. 9, 2010, 44-48.*

vii. *S.Sivakumar and Dr.C.Chandrasekar," Modified PSO Based Feature Selection for Classification of Lung CT Images", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5, No.2, 2014, Pp: 2095-2098.*

viii. *S.Sivakumar and Dr.C.Chandrasekar," Feature Selection using Genetic Algorithm with Mutual Information", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5, No.3, 2014, Pp: 2871-2874.*

ix. *V. Tereshko and T. Lee. How information mapping patterns determine foraging behavior of a honey bee colony. Open Systems and Information Dynamics, 9:181-193, 2002.*

x. *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.*