International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-1, January, 2019

Filter based Feature Selection using ABC

Dr. S. Sivakumar
Department of Computer Science
Periyar University
Salem, India
E-mail: ssivakkumarr@yahoo.com

Abstract: Classification is an important task in machine learning and data mining, which aims to classify each instance in the data into different groups. The feature space of a classification problem is a key factor influencing the performance of a classification/learning algorithm. Without prior knowledge, it's hard to determine which features are useful. Therefore, a large number of features are usually introduced into the dataset, including relevant, irrelevant and redundant features. However, irrelevant and redundant features are not useful for classification. Their presence may mask or obscure the useful information provided by relevant features, and hence reduces the quality of the whole feature set. Meanwhile, the large number of features causes one of the major obstacles in classification known as "the curse of dimensionality". Therefore, feature selection is proposed to increase the quality of the feature space, reduce the number of features and improve the classification performance. In this paper Artificial Bees Colony (ABC) algorithm is used to evaluate the feature selection performance with mutual information on different UCI Data Repository datasets.

Keywords: feature selection, artificial bees colony, mutual information, classification.

I. INTRODUCTION

Feature selection aims to select a subset of relevant features that are necessary and sufficient to describe the target concept [1]. By reducing the irrelevant and redundant features, feature selection could decrease the dimensionality, reduce the amount of data needed for the learning process, shorten the running time, simplify the structure and/or improve the performance of the learnt classifiers [1].

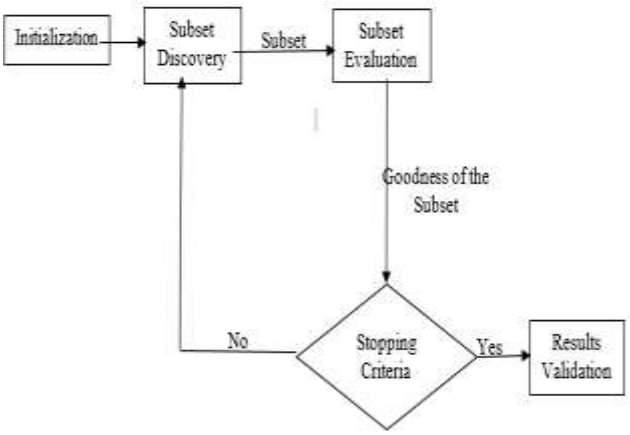

Figure 1: Feature subset selection and evaluation process

Naturally, an optimal feature subset is the smallest feature subset that can obtain the optimal performance, which makes feature selection a multi-objective problem [2]. Note that feature selection algorithms choose a subset of features from the original feature set and do not create new features. Feature selection is a difficult task. Although many approaches have been proposed, most of them still suffer from the problems of stagnation in local optima and high computational cost due mainly to the large search space. Therefore, an efficient global search technique is needed to address feature selection tasks.

II. CHALLENGES IN FEATURE SELECTION APPROACHES

Feature selection is a difficult problem [3,4], especially when the number of available features is large. The task is challenging due mainly to two reasons, which are feature interaction and the large search space. Feature interaction (also called epistasis [5]) frequently happens in classification tasks. There can be two-way, three-way or complex multiway interactions among features. On one hand, a feature, which is weakly relevant or even entirely irrelevant to the target concept by itself, can significantly improve the classification accuracy if it is complementary to other features. Therefore, the removal of such features may also miss the optimal feature subsets. On the other hand, an individually relevant feature can become redundant when working together with other features. The selection/use of such features brings redundancy, which may deteriorate the classification performance. In feature selection, the size of the search space grows exponentially with respect to the number of available features in the dataset (2^n possible subsets for n features) [6]. In most cases, it is practically impossible to search exhaustively all the candidate solutions. To better address this problem, a variety of search techniques have been applied to feature selection [1, 6]. However, existing methods still suffer from the problem of stagnation in local optima and/or high computational cost.

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 goals are usually conflicting to each other, and the optimal decision needs to be made in the presence of a trade-off between them. Treating feature selection as a multi-objective problem can obtain a set of non-dominated feature subsets to meet different requirements in real-world applications. However, there are rare studies treating feature selection as a multi-objective problem [7, 8].

Two key factors in a feature selection algorithm are the search strategy and the evaluation criterion. The search

space of a feature selection problem has 2^n possible points/solutions, where n is the number of available features. The algorithm explores the search space of different feature combinations to find the best feature subset. However, the size of the search space is huge, especially when the number of features is large. This is one of the main reasons making feature selection a challenging task.

III. FEATURE SELECTION APPROACHES

Existing feature selection methods can be broadly classified into two categories: filter approaches and wrapper approaches. Wrapper methods include a classification algorithm as a part of the evaluation function to determine the goodness of the selected feature subsets. Filter methods use statistical characteristics of the data for evaluation, and the feature selection search process is independent of any classification algorithm. Filter methods are computationally less expensive and more general than wrapper procedures while wrappers are better than filters in terms of the classification performance [1].

Wrapper based Feature Selection:

In a wrapper model, the feature selection algorithm exists as a wrapper around a classification algorithm and the classification algorithm is used as a "black box" by the feature selection algorithm [9]. The performance of the classification algorithm is employed in the evaluation function to evaluate the goodness of feature subsets and guide the search.

Filter based Feature Selection:

In filter algorithms, the search process is independent of any classification algorithm. The goodness of feature subsets are evaluated based on a particular criterion like distance measure, information measure and consistency measure [1]. Filter algorithms are argued to be computationally less expensive and more general than wrapper algorithms [9, 10], but filter algorithms totally ignore the performance of the selected feature subset on the classification algorithm, which usually leads to lower performance than wrapper algorithms on a particular classification algorithm [9]. Compared with filter algorithms, wrappers often produce better classification performance because of the interaction between the classification algorithm and the selected feature subsets during the feature selection process [11]. However, wrapper feature selection algorithms are usually computationally more expensive than filters because each evaluation of a candidate solution needs a learning/classification algorithm to be trained and tested [10].

IV. MUTUAL INFORMATION FOR FEATURE SELECTION

When there are thousands of features, wrapper approaches become infeasible because the evaluation of large feature sets is computationally expensive. Filter methods evaluate feature subsets via different statistical measures. Among the filter approaches, a fast way to assess individual features is given by their relevance to the classification, by maximizing the mutual information between each single variable and the classification output. In this work we use the mutual information criterion and we estimate its value directly from the data. This kind of estimation methods bypasses the estimate of the distribution of the samples. Thus, the small number of samples in a high dimensionality is not a problem anymore. Information theory offers a solid theoretical framework for many different machine learning problems. Information theory [12,13] provides a solid theoretical framework for many different machine learning problems. In the case of feature selection, information theoretic methods are usually applied in the filter feature selection way. A classical use of information theory is found in several feature ranking measures. These consist of statistics from the data that score each feature F_i depending on its relation to the classes C_i.

$$I(F,C) = \int\int p(f,c) log \frac{p(f,c)}{p(f)p(c)} df dc \qquad (1)$$

Some approaches evaluate the mutual information between a single feature and the class label. This measure is not a problem. The difficulties arise when evaluating entire feature sets. The necessity for evaluating entire feature sets in a multivariate way is due to the possible interactions among features. While two single features might not provide enough information about the class, the combination of both of them could, in some cases, provide significant information. In they show experimentally that, as $I(F_1; F_2; C)$ decreases, higher is the need of a multivariate model [13]. While two single features might not provide enough information about the class, the combination of both of them could, in some cases, provide significant information. For the mutual information between N variables $X_1, X_2 \ldots X_N$, and the variable Y, the chain rule is [13]:

$$I(X_1, X_2, X_3, \ldots..X_N; Y) = \sum_{i=1}^{N} I(Xi; Y|Xi-1, Xi-2, \ldots..., X1) \qquad (2)$$

The usual approach for calculating mutual information is to measure entropy and substitute it in the mutual information formula. Mutual information is considered to be a suitable criterion for feature selection. Mutual information is a measure of the reduction of uncertainty about the class labels, due to the knowledge of the features of a data set.

V. ARTIFICIAL BEES COLONY OPTIMIZATION (ABC)

The artificial bee colony (ABC) algorithm is a new population-based metaheuristic technique based on the foraging behavior of honey bee swarm. Karaboga initially developed the ABC algorithm in 2005. On the basis of their foraging behavior, real bees are divided into three categories namely employed, scouts and onlookers. Employed bees are those bees that are currently exploiting the food sources [14]. 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.

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.

The Basic ABC algorithm
Initialize Population Repeat Place the employed bees on their food sources Place the onlooker bees on the food sources depending on their nectar amounts Send the scouts to the search area for discovering new food sources Memorize the best food source found so far Until requirements are met

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.

Steps involved in the ABC algorithm:

The following steps are involved in the ABC algorithm.

1. Initialize the population of solutions x
2. Evaluate the population
3. Cycle = 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

$$(V_{i,j}=X_{i,j} + \Phi_{i,j} (X_{i,j}-X_{k,j}) \quad (3)$$

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 values using (4).

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \quad (4)$$

In order to calculate the fitness values of solutions using (5) or (6):

$$fit_i = \left\{ \frac{1}{1+f_i} \quad if\ f_i \geq 0 \right\} \quad (5)$$

$$fit_i = \left\{ 1 + abs(f_i) \quad if\ f_i < 0 \right\} \quad (6)$$

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 (7).

$$X_{i,j} = min_j + rand(0,1)*(max_j-min_j) \quad (7)$$

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

The quality of the solution represented by that 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 concerned with 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.

VI. DATASETS

In this paper, the following UCI Machine Learning Datasets are used to evaluate the performance the feature selection algorithm with classification accuracy.

Table 1: List of Datasets

Name of the Dataset	Number of Instances	Number of Attributes with class
Iris	150	4
Lung Cancer	32	56
SPECTF Heart	267	44
Wine	178	13
Yeast	1484	8

VII. PERFORMANCE ANALYSIS

The performance of the Mutual Information based feature selection is evaluated with k-NN classifier with 10-fold cross validation. The following table shows the original dataset classification accuracy and the feature selection applied dataset accuracy with selected features.

Table 2: Classification Accuracy of Original and Selected Features

Name of the Dataset	Number of attributes	Original dataset classification accuracy (%)	Number of Selected attributes with ABC	Classification accuracy of selected attributes (%)
Iris	3	72	2	77
Lung Cancer	55	67	31	94
SPECTF Heart	43	73	28	91
Wine	13	81	8	93
Yeast	7	84	5	95

From the Table2, the Mutual Information based feature selection with ABC optimization technique which reduces more number of subsets and increases classification accuracy over the given original subset.

VIII. CONCLUSION

In this paper, ABC based feature selection algorithms with mutual information works well for various UCI Machine Learning Datasets. The results show that the minimized subset produces a good classification results over the original feature subset. In order to improve the classification accuracy, we need a much relevant feature subset.

REFERENCES

[1] S. J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", Second Edition, Pearson Education, 2003.

[2] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, Vol. 97, pp. 273–324, 1997.

[3] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Feature selection using multi-objective genetic algorithms for handwritten digit recognition," in 16th International Conference on Pattern Recognition (ICPR'02), Vol. 1, pp. 568– 571, 2002.

[4] C. 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] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems" Theoretical Computer Science, Vol. 209, pp. 237–260, 1998.

[6] M. Mitchell, "An Introduction to Genetic Algorithms", The MIT Press, 1996.

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", The Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[8] 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.

[9] 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.

[10] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem" in Machine Learning: Proceedings of the Eleventh International Conference (ICCCS'11), Morgan Kaufmann Publishers, pp. 121–129, 1994.

[11] C. S. Yang, L. Y. Chuang, and J. C. Li, "Chaotic maps in binary particle swarm optimization for feature selection", in IEEE Conference on Soft Computing in Industrial Applications (SMCIA '08), pp. 107–112, 2008.

[12] M. Cover and J. Thomas, "Elements of Information Theory", Wiley Interscience, 1991.

[13] A. P. Mart´ınez, P. Larra˜naga, and I. Inza, "Information theory and classification error in probabilistic classifiers", In Discovery Science, pages 347–351, 2006.

[14] D. Karaboga and 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.

AUTHOR'S BIOGRAPHY

Dr. S. Sivakumar: Presently working as Teaching Assistant in Department of Computer Science, Periyar University, Salem, TN (India). He received the Ph.D. degree in Computer Science from Periyar University. He has published several research papers in National and international journals in his credit.