I.J. Intelligent Systems and Applications, **2016, 3, 42-47**
Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2016.03.05

IGICA: A Hybrid Feature Selection Approach in Text Categorization

Mohammad Mojaveriyan
Department of Computer and Electrical Engineering, University of Kashan, Kashan, Iran
E-mail: Mojaveriyan.kh@gmail.com

Hossein Ebrahimpour-komleh
Department of Computer and Electrical Engineering, University of Kashan, Iran
E-mail: ebrahimpour@kashanu.ac.ir,

Seyed jalaleddin Mousavirad*
Department of Computer and Electrical Engineering, University of Kashan, Iran
E-mail: jalalmoosavirad@gmail.com, s.mousavirad.2015@ieee.org

***Abstract*—**Feature selection problem is one of the most important issues in machine learning and statistical pattern recognition. This problem is important in many applications such as text categorization because there are many redundant and irrelevant features in these applications which may reduce the classification performance. Indeed, feature selection is a method to select an appropriate subset of features for increasing the performance of learning algorithms. In the text categorization, there are many features which most of them are redundant. In this paper, a two-stage feature selection method-IGICA- based on imperialist competitive algorithm (ICA) is proposed. ICA is a new metaheuristic which is inspired by imperialist competition among countries. At the first stage of the proposed algorithm, a filtering technique using the information gain is applied and features are ranked based on their values. The top ranking features are then selected. In the second stage, ICA is applied to the select the efficient features. The presented method is evaluated on Retures-21578 dataset. The experimental results showed that the proposed method has a good ability to select efficient features compared to other methods.

Index Terms—Text classification, Feature selection, Imperialist competition algorithm, Information gain.

I. Introduction

Feature selection problem in many applications is important because there are a lot of redundant, irrelevant, and repetitive features in these applications. Removing these features lead to acceleration of the learning algorithm and the improvement of its performance. The high dimensional feature space is one of the main problems in text categorization which can be decreased the efficiency of learning algorithm because of the redundant or irrelevant features. It is essential to select a suitable subset of features to increase the performance of categorization. There are many algorithms for feature selection problem. These methods have attempt to find the best subset from 2N candidate subset. These methods are divided into two general groups of filter and wrapper. Filter method utilizes a criterion for assigning weights to the features and then feature selection is performed according to the weights [1]. The computational cost of these methods is low and the performance depends on the quality of weighting criteria. Features are selected based on a learning algorithm in wrapper methods. Each of these groups has various methods which a few ones are used for large text categorization problems [2]. It has been shown that most of these methods are not suitable for high dimensional feature spaces[1,2].

In this paper, a two-stage approach based on ICA, IGICA, is presented for feature selection problem that originated from Reuter. Our main work in this is on obtaining the best Accuracy, recall and F-measure. At the first stage, a filtering technique called information gain is applied and features are ranked based on their values. Features with the highest ranking are then selected. In the second stage, ICA is applied to the select of remaining features. Imperialist competitive algorithm (ICA) is a new metaheuristic which is inspired from imperialist competition among countries. This algorithm was introduced by Atashpaz-Gargari and Caros-Lucas[7]. This algorithm begins with a number of countries (which are "solutions") in the original form. Countries are divided into two groups of imperialist and colony countries. After creating the empire, imperialist countries start to improve their colonies. In fact, in this algorithm the most powerful empires try increase their power by possession of colonies while weak empires collapse so powerful empires try to take possession of colonies of other empires. This algorithm is applied for feature selection problem [8,9] and has been able to demonstrate a good performance but its performance on the text feature selection has not been investigated. This paper is the first attempt to the text feature selection problem. The next part of this paper describes the feature selection problem and the next part describes the proposed method. Finally, the experimental results and conclusion are considered.

II. Related work

Feature selection is process which in that a subset of the features based on criteria are selected. By now, lots of achievements after first paper in the case of feature selection have been obtained. Feature selection in various fields, in order to forecast, classification, clustering and reduced the size of data.

Several methods have been presented to select features in the text categorization. Yang and Pedersen [2] compared several different feature selection algorithms including document frequency, information gain, mutual information, chi-square, and the power of words. The results showed that information gain and chi-square criteria are effective to select a subset of features without reducing of classification accuracy. In another work, Chen et al. [3] are evaluated two different criteria for feature selection on text dataset, multi-class odds ratio and class discriminating measure. Results showed that these two measures produce more valuable results. George comparison of twelve feature selection methods(for example information gain, chi, odds Ratio, probability Ratio, Document Frequency, ect) evaluated on a dataset REUTERS, TEREC, OHSUMED, ect and he present a new feature selection method named 'Bi-Normal Separation'[4]. Tehseen and et al in[5] evaluation of feature selection(information gain, chi and symmetrical uncertain) for Urdu text classification and they are using six classification algorithm(na ïve bays, k-nearest neighbor, support vector machines, polynomial and radial basis kernels and decision tree).

Metaheuristic approaches such as ant colony, genetic algorithm, and particle swarm optimization have been developed for different optimization problems. These algorithms have attracted a lot of attention in text feature selection in recent years. Fardin Ahmadizar et. al [6] developed a two-phase feature selection strategy.In the first phase, a number of features are removed based on the fuzzy entropy and in the second phase, the remaining features are limited using the ant algorithm. Aghdam et al[7] is used the ant colony algorithm to select the appropriate features. In another work, Harun[8] proposed a two-stage feature selection method for text categorization by using information gain, principal component analysis(PCA) and genetic algorithm. PCA is the technique used to produce the lower-dimensional the original dataset. In the first phase, a number of features are removed based on the information gain and in the second phase, the remaining features are limited using the genetic algorithm. They have used to evaluate the effectiveness of their purposed model, experiments are conducted using the k-nearest neighbor and C4.5 decision tree algorithm and they have used on Reuters-21,578 and Classic3 datasets collection for text categorization.

III. Feature Selection

Critical problems in text categorization is the high dimensional feature space. In fact, the speed and accuracy of categorization is reduced when there are many redundant and irrelevant features. Feature selection can be considered as an effective tool to identify irrelevant and redundant features [9,12]. Feature selection problem in text categorization is an important issue because it will help to eliminate redundant features with little information and to improve the categorization performance.

Feature selection methods are divided into two general groups: 1-filter 2-wrapper. In the filter method, unlike the wrappers, feature selection is independent of learning algorithm. In other words, there is no feedback from the learning algorithm in the filter methods. First, a validation rank is calculated for every feature in filter methods, and features with the lowest validation rank are then removed from the feature set. Wrapper methods use a learning algorithm to evaluate the feature set. Thus the features are selected based on their effect on the accuracy improvement.

IV. The Proposed Feature Selection Algorithm

A two-stage feature selection method is presented in this paper for text categorization. In the first phase, redundant and irrelevant features are omitted using information gain method (which is a filter strategy) and the remaining features are transferred into the second phase. In the first stage, features are independently considered. In the next stage, ICA algorithm is used. Indeed, in this stage, dependencies between features is also examined. Finally, the nearest neighbor algorithm is applied for categorization.

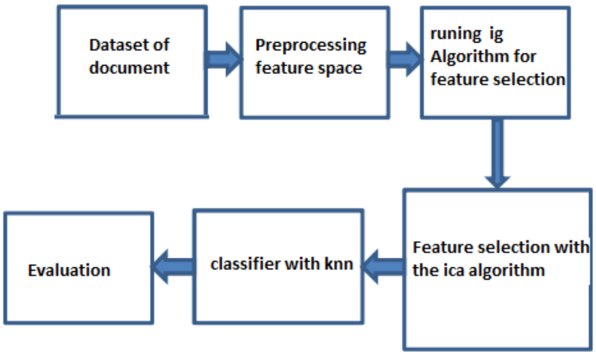

Dataset of document
Preprocessing feature space
runing ig Algorithm for feature selection
Feature selection with the ica algorithm
classifier with knn
Evaluation

Fig.1. Steps in the proposed model for text classification using feature selection.

A. *Information Gain*

Information gain (IG) is successively used in machine learning methods as the criterion of feature quality. It is defined as follows:

$$\begin{aligned} G(t) = & -\sum_{i=1}^{m} p_r(c_i)\log p_r(c_i) \\ & + p_r(t)\sum_{i=1}^{m} p_r(c_i|t)\log p_r(c_i|t) \\ & + p_r(\bar{t})\sum_{i=1}^{m} p_r(c_i\,|\,\bar{t})\log p_r(c_i\,|\,\bar{t}). \end{aligned} \quad (1)$$

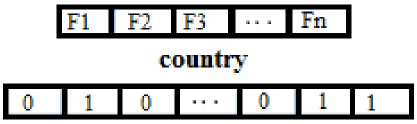
Where pr(ci|t) is the probability value which in a document x of the training set the feature t appears and the document x belongs to the class c. Also $p(c|\bar{t})$ shows the probability value that in a document x of the training set the feature t does not appears and the document x belongs to the class c. values of $p(\bar{t})$, p(t) , p(c) show no occurrence of feature, probability of occurrence of feature, and the probability of occurrence of class in document x, respectively.

In the first stage, for each word t in each class c, this value is calculated and finally the maximum of these values is considered as the information gain of a word. Then their highest is selected as the best features in the first stage.

B. *Imperialist Competitive Algorithm*

Imperialist competitive algorithm (ICA) is a new metaheuristic that has been recently introduced for solving different optimization problems. This method is inspired by imperialist competitive among countries [7]. Colonization is a historical inevitable phenomenon in which some powerful countries (military and economic power) have colonized some weaker countries in order to obtain their resources. This method like other evolutionary algorithms, begins with a primary random population called country which are the problem solutions. Countries are then divided into two groups: imperialist and colony countries of which a number of countries with the best criteria (such as best accuracy in categorization) are selected as the imperialists and the others are selected as the colonies of these imperialists. Running this algorithm will simulate a kind of competition between imperialists.

Similar other evolutionary algorithms, this algorithm starts by generating a set of candidate random solutions. The generated random points are called the country. Countries in this algorithm are such as chromosomes in genetic algorithm and particles in particle swarm optimization. In the IGICA, a country is a 1×N binary array. When value of a feature is 1, then the feature is selected and but if it is 0, the feature is not selected. Figure1 represents a typical country.



F1	F2	F3	···	Fn

country

0	1	0	···	0	1	1

Fig.2. A subset of features in the IGICA (country)

The cost value of a country is determined by classification accuracy. Many classifiers can be used to compute the classification accuracy such as KNN, decision tree and support sector machine.We used KNN classifier to obtain the cost function because KNN is simple and quick. Classification accuracy is defined as follows:

$$cost_{function} = \frac{Number\ of\ correctly\ classified\ samples}{Total\ Number\ of\ samples} \quad (2)$$

Then for example 10 percent of countries which their cost function is higher than the rest of the countries are selected to be the imperialist states and the rest of them form the colonies of these imperialists. Then Imperialist competition algorithm is performed and an empire is collapse if it lose all colonies. In this competition, all empires try to take ownership of colonies of other empires. Figure 2 shows a modeled imperialist competition.

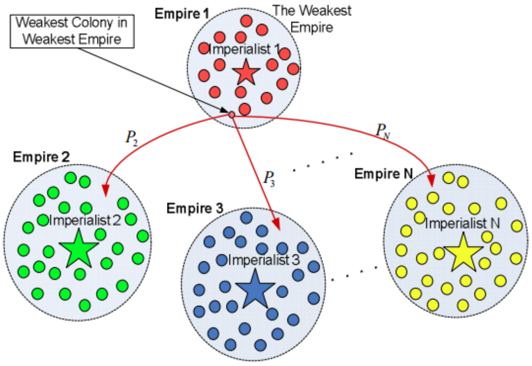


Fig.3. Imperialist competition [9].

Two main operators of this algorithm are assimilation and revolution. These operators are described as follows:

1- Assimilation

After the initial organization of the empires, colonies in each imperialist get closer to the related imperialist country. This displacement is a simple model of assimilation policy. Since the feature selection problem is discrete, in the current research, a new assimilation [11] is used which is suitable for discrete problems. It is presented as follows

For colonies and each imperialist:

1) Create a randomly binary array to each imperialist and their colonies
2) Copy the cells from the imperialist corresponding to the location of the 1s in the binary array to the same positions in the colonies.

2- Revolution

Revolution brings about sudden random changes in the position of some countries in the search space. This policy is similar to mutation process in genetic algorithm (GA). The revolution increases the exploration of the algorithm and prevents the early convergence. In ICA, the revolution rate defines the percentage of colonies in each empire that undergoes the revolution process. In each iteration, for every colony, a random number which is varying between 0 and 1 is generated. Then, this value is compared with the probability of revolution rate. If random number is lower than probability revolution rate, the revolution is performed. The new colony will replace with previous colony while its cost is improved [11].

It is clear that the power of an empire depends on both the power of the imperialist country and the power of its colonies. Moreover, the power of imperialist has greater impact on the total power of an empire while colonies power has lower effect. This is modeled by defining the total cost of an empire:

$$\begin{aligned} T_{Cn} &= cost\{imperialistn\}C \\ &+\xi\ mean\{cost(colonies\ of\ empiren)\} \end{aligned} \tag{3}$$

where TCn is the total cost of the nth empire and ξ is a positive number which is weighted to be less than 1.

Based on their total power, in this competition, each of empires will have a probability of taking ownership of the said colonies. In other words, these colonies will not be attracted by the strongest empires, but these empires will be more probably to own them. When an empire miss all of its colonies, it is supposed to be crumbled. After a while, all the empires will crumble and all the colonies will be under the control of this unique empire. In this ideal model, all colonies have the same position and same costs and they are controlled by an imperialist, which means the algorithm converges to the best solution.

The above steps of imperialist competition based feature selection can be summarized as the below pseudo code:

```
1) Generate some random binary array in the search space and create initial empires.
2) Assimilation: Colonies move towards their relevant imperialist states.
3) Revolution: Random changes occur in the characteristics of some countries.
4) Position exchange between a colony and imperialist. A colony with a better position than the imperialist has the more chances to take the control of empire by replacing the existing imperialist.
5) Imperialist competition: All imperialists compete to take owner ship of colonies of each other.
6) Eliminate the powerless empires. Weak empires lose their power gradually and they will finally be eliminated.
7) If the stop condition is satisfied, stop, if not go to 2.
```

V. Experimental Results and Analysis

In this section, first the used dataset is described. Then, the feature extraction method in text categorization is introduced. Finally, the computational results of the proposed method for feature selection is presented.

A. The used Dataset

There are many standard datasets to text categorization such as reuters-21578.We have applied the Reuters-21578 to evaluate the proposed method because of its popular. This dataset is belonged to the Reuters news in 1987 which includes 21578 textual documents in different groups. We have applied top six classes among 7285 documents which are divided in to two groups, 5228 documents for training set and 2057 documents for test set.

B. Feature extraction

In the feature extraction process, document (di) is as a vector that is represented the weight of a word in document j as follows:

$$dj=\{W1j,\ W2j,\ldots..,W|r|j\}$$

In this formula , wij represents the weight of the ith word in the jth document.

Each term vector indicates the occurrence of a word in a document. These weights can be normalized by the tfidf function .This function shows a numerical value to present the importance of a word in a document[15].

C. Performance Evaluation

In order to evaluate the text categorization, three measures precision, recall and the F-measure are usually used. The precision represents the categorization accuracy and recall is a measure of completeness. The higher level of these values is indicated better quality of categorization method. The calculations of precision and recall are mentioned in below:

$$precision = \frac{fp}{tp + fp} \tag{4}$$

$$Reacall = \frac{fp}{tp + fn} \tag{5}$$

which TPi, FPi, FNi are true-positive, false positive and false negative. Therefore, precision is defined as the ratio of the documents that are classified correctly to the total documents, recall determines the ratio of the documents that are classified correctly to the test data. One other categorization measures is F-measure which is a combination of recall and precision measures [16] .It is balanced combination of precision and recall.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

In the text categorization, there are two popular methods for calculating performance based on precision and recall. These two methods are called micro-averaging and macro-averaging. The significant difference between them is that the micro-averaging gives an equal value to the all documents so it is document- based while the macro-averaging gives an equal value to each class so it is classification-based.

The formulas 7 to10 show the micro and macro averaging on both the precision and the recall:

$$precision_{micro} = \frac{\Sigma_{j=1}^{|c|} TPj}{\Sigma_{j=1}^{|c|} \left(TPj + FPj\right)} \qquad (7)$$

$$recall_{micro} = \frac{\Sigma_{j=1}^{|c|} TPj}{\Sigma_{j=1}^{|c|} \left(TPj + FNj\right)} \qquad (8)$$

$$precision_{macro} = \frac{\sum_{j=1}^{|c|} {}_{percision}\, j}{|c|} \qquad (9)$$

$$recall_{macro} = \frac{\sum_{j=1}^{|c|} {}_{recall}\, j}{|c|} \qquad (10)$$

D. Evaluation of the Proposed Method

In order to adjust the parameters of imperialist competition algorithm, several experiments have carried out in a trial and error manner. Best value of parameters is shown in table (1). Although, it is possible that the better parameters is achieved with more experiments.

Table 1. The Parameters Values Used For Selecting Features

Parameter	Value
Society number	200
Number of emperors	15
Maximum of decades	200
Revolution rate	0.0001
Zeta-value	0.001

To evaluate the results, at the first level M percent of extracted features are selected based on the information gain method (m is the percentage of features for example 5%, 7%, 12.5%, 15%). Then in the second stage of the algorithm, an ICA based feature selection method is used in order to select the efficient features from m features of the previous stage. The results of the information gain and the proposed algorithm is shown in table 2 and 3. As can be seen, feature selection with 15% features is presented the highest performance using information gain while the IGICA method is investigated it with 12% features.

Table 4 compares the proposed method with other algorithms. According to the results, the proposed method has a higher F-measure compared to the other methods.

Table 2.Classification with the Information Gain

M	Precision	Recall	F-measure	F1-micro
5	88.40	64.59	74.64	88.40
7.5	89.12	66.30	76.03	89.12
12	89.17	68.30	76.034	89.17
15	89.37	70.46	78.79	89.37

Table 3. Classification with the proposed method (M is the percentage of features in the first stage)

M	Precision	Recall	F-measure	F1-micro
5	94.53	86.48	90.32	94.53
7.5	94.84	86.74	90.60	94.84
12	95.20	88.01	91.46	95.20
15	94.94	87.25	90.932	94.94

Table 4. Comparison of The existing method with the proposed method (M=5%)

Method name	Precision	Recall	F-measure
Information gain	84.72	65.13	73.64
Mutual information	53.14	50.00	51.52
The proposed method	94.53	86.48	90.32

The advantage of the proposed method is combination wrapper technique and filter technique. In fact, in the first stage, a filter method is applied. In this stage, some features are independently of other features are selected. In the second part of the proposed algorithm, a wrapper method to remove features are discussed.

VI. Conclusion

In this article, a two-stage feature selection method, IGICA, was presented. The results showed that the proposed feature selection approach increases the efficiency of text categorization. In the first phase, information gain is applied to eliminate the redundant and irrelevant features and the remaining features will be transferred to the second stage. Imperialist competitive algorithm is used for feature selection. The main advantage of the proposed approach is combination of filter and wrapper methods. In order to classify the documents, k-nearest neighbor classifier is used. Results showed that the proposed approach can increase the efficiency of document categorization compared to the compared methods.

References

[1] Yang, Y. and Pedersen, J. A. (1997). “A comparative study on feature selection in text categorization.” In Proceedings of 14th International Conference on Machine Learning (ICML-97), PP.412-420.

[2] Y. Yang, and J. O. Pedersen, "A comparative study on feature selection in text categorization," In Proceedings of the 14th International Conference on Machine Learning, pp. 412-420,1997.

[3] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature Selection for text classification with Naive Bayes," Expert SystAppl, vol. 36, pp. 5432-5435,2009.

[4] Forman, George. "An extensive empirical study of feature selection metrics for text classification." The Journal of machine learning research 3 (2003): 1289-1305.

[5] Zia, Tehseen, Qaiser Abbas, and Muhammad Pervez Akhtar. "Evaluation of Feature Selection Approaches for Urdu Text Categorization." (2015).

[6] Ahmadizar, Fardin, Majid Hemmati, and Ahmad

Rabanimotlagh. "Two-stage text feature selection method using fuzzy entropy measure and an t colony optimization." Electrical Engineering (ICEE), 2012 20th Iranian Conference on. IEEE, 2012.

[7] Aghdam, Mehdi Hosseinzadeh, Nasser Ghasem-Aghaee, and Mohammad EhsanBasiri. "Text feature selection using ant colony optimization." Expert systems with applications 36.3 (2009): 6843-6853.

[8] Uğuz, Harun. "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm." Knowledge-Based Systems 24.7 (2011): 1024-1032.

[9] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition," in Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 4661- 4667.

[10] Mousavirad, S. J., and H. Ebrahimpour-Komleh. "Feature selection using modified imperialist competitive algorithm." Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on. IEEE, 2013.

[11] SJ Mousavirad, F. Akhlaghian Tab, and K. Mollazade. "Application of imperialist competitive algorithm for feature selection: A case study on bulk rice classification." International Journal of Computer Applications (0975–8887) Volume (2012).

[12] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," J Mach Learn Res, vol. 3, pp. 1157-1182, 2003.

[13] H. Liu, and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE T Knowl Data En, vol. 17, iss. 4, pp. 491-502, 2005.

[14] The Reuters -21578 text categorization test Collection.http://kdd.ics.uci.edu/databaseslreuters21578/reuters21578.html

[15] G. Salton, and C. Buckley, Term-weighting approaches in automatic text retrieval. Cornell University Ithaca: NY, TR87-881, 1987.

[16] C. 1. van Rijsbergen, Information Retrieval, 2nd ed., Butterworth:London, 1979.

[17] Ahmadizar, Fardin, Majid Hemmati, and Ahmad Rabanimotlagh. "Two-stage text feature selection method using fuzzy entropy measure and an t colony optimization." Electrical Engineering (ICEE), 2012 20th Iranian Conference on. IEEE, 2012.

Authors' Profiles

Mohamad mojaveriyan was born in mashhad, Iran, in 1988. He graduated from khayam University of Mashhad in 2011. He started his MA degree at Computer engineering department in the University of Kashan, Iran, in 2012. His main research interests are Data Mining, Text mining, Feature selection, distributed systems, and Combination systems. He has published some papers in international conferences and journals. Presently, He is working in the area of feature selection in text.

Hossein Ebrahimpour-Komleh is currently an Assistant Professor at the Department of Electrical and Computer Engineering at the University of Kashan, Kashan, Iran. His main area of research includes Computer vision, Image Processing, Pattern Recognition, Biometrics, Robotics, Fractals, chaos theory and applications of Artificial Intelligence in Engineering. He received his Ph.D. degree in Computer engineering from Queensland University of technology, Brisbane, Australia in 2006. His Ph.D. research work was on the “Fractal Techniques for face recognition”. From 2005 to 2007 and prior to joining the University of Kashan, he was working as a Post-doc researcher in the University of Newcastle, NSW, Australia and as a visiting scientist in CSRIO Sydney. Hossein Ebrahimpour-Komleh has B.Sc. and M.Sc. degrees both in computer engineering from Isfahan University of Technology (Isfahan, Iran) and Amirkabir University of Technology (Tehran, Iran,) respectively. He has served as the editorial board member and reviewer of several journals and international and national conferences.

Seyed Jalaleddin Mousavirad was born in Neishabour, Iran, in 1986. He graduated from Azad University of Mashhad in 2007. He received MA degree from Kurdistan University, Iran in 2011 & started his Ph.D. at Computer engineering Department in the University of Kashan, Iran, in 2012. His main research interests are metaheuristic algorithms, Data Mining, Pattern Recognition, and Image processing. He has published about 15 research papers in National and International journals and conferences. Presently, He is working in the area of metaheuristic algorithms.