IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 6, Ver. VI (Nov – Dec. 2015), PP 45-51
www.iosrjournals.org

Filter-Wrapper Approach to Feature Selection Using PSO-GA for Arabic Document Classification with Naive Bayes Multinomial

Indriyani[1)], Wawan Gunawan [2)], Ardhon Rakhmadi[3)]
[1, 2, 3)] *Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia, Sukolilo, Surabaya, 60117, Indonesia*

Abstract: *Text categorization and feature selection are two of the many text data mining problems. In text categorization, the document that contains a collection of text will be changed to the dataset format, the dataset that consists of features and class, words become features and categories of documents become class on this dataset. The number of features that too many can cause a decrease in performance of classifier because many of the features that are redundant and not optimal so that feature selection is required to select the optimal features. This paper proposed a feature selection strategy based on Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) methods for Arabic Document Classification with Naive Bayes Multinomial (NBM). Particle Swarm Optimization (PSO) is adopted in the first phase with the aim to eliminate the insignificant features and prepared the reduce features to the next phase. In the second phase, the reduced features are optimized using the new evolutionary computation method, Genetic Algorithm (GA). These methods have greatly reduced the features and achieved higher classification compared with full features without features selection. From the experiment that has been done the obtained results of accuracy are NBM 85.31%, NBM-PSO 83.91% and NBM-PSO-GA 90.20%.*

Keywords: *Document Classification, Feature Selection, Particle Swarm Optimization (PSO), Genetic Algorithm (GA).*

I. Introduction

Text data mining[1] is a research domain involving many research areas, such as natural language processing, machine learning, information retrieval[2], and data mining. Text categorization and feature selection are two of the many text data mining problems. The text document categorization problem has been studied by many researchers[3][4][5]. Yang et al. showed comparative research on feature selection in text classification[5]. Many machine learning methods have been used for the text categorization problem, but the problem of feature selection is to find a subset of features for optimal classification. Even some noise features may sharply reduce the classification accuracy. Furthermore, a high number of features can slow down the classification process or even make some classifiers inapplicable.

According to John, Kohavi, and Pfleger[6], there are mainly two types of feature selection methods in machine learning: wrappers and filters. Wrappers use the classification accuracy of some learning algorithm as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification.

Hence, feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers. Text classification tasks can usually be accurately performed with less than 100 words in simple cases, and do best with words in the thousands in the complex ones[7].

As opposed to wrappers, filters perform feature selection independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class.

Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification.

For this reason, our motivation is to build a good text classifier by investigating hybrid filter and wrapper method for feature selection. Finally, our proposed algorithm Particle Swarm Optimation

for filtering and second phase, we reduce result features filtering using wrapper Genetic Algorithm (GA) for optimizing features for Classification using Naive Bayes Multinominal for Arabic Document text Classification. The evaluation used an Arabic corpus that consists of 478 documents from www.shamela.ws, which are independently classified into seven categories.

II. Methodology

General description of the research method is shown in Figure 1. The stages and the methods used to lead this study include.

Stages and methods used to underpin this study include:

A. DataSet Document

Document used as experimental data taken from www.shamela.ws and taken seven categories according to the most often discussed is Sholat, Zakat, Shaum, hajj, mutazawwij, Baa'aand Isytaro and Wakaf/

B. PreProcessing

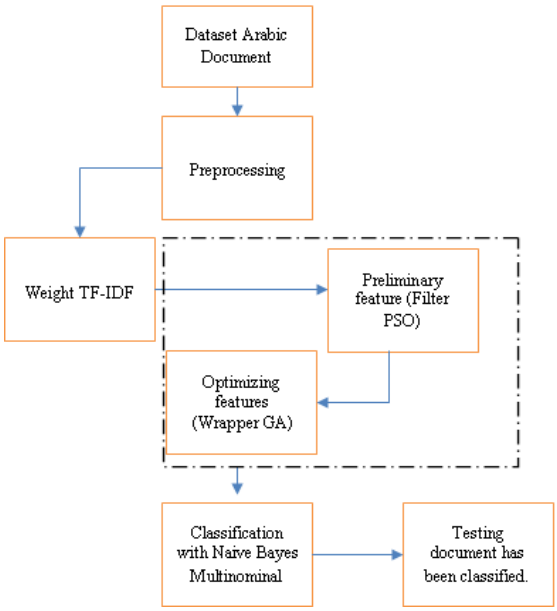

Fig 1. Research Stage

In the process of preprocessing, each document is converted to vector document news with the following sequence:

- Filtering, namely the elimination of illegal characters (numbers and symbols) in the document's contents.
- Stoplist removal, i.e. removal of the characters included in the category of stopword or words that have a high frequency contained in the data stoplist.
- Terms extraction, which extract the terms (words) of each document to be processed and compiled into a vector of terms that represent the document.
- Stemming, namely to restore the basic shape of each term found in the document vector and grouping based on the terms similar.
- TF-IDF weighting i.e. performs weighting TF-IDF on any terms that exist in the document vector.

C. Filter - Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart [11]. The Particle Swarms find optimal regions of the complex search space through the interaction of individuals in the population. PSO is attractive for feature selection in that particle swarms will discover best feature combinations as they fly within the subset space.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*) and it must be in a range defined by parameters *vmin* and *vmax*. The personal best position of the particle *local best* is the best previous position of that particle and the best position obtained by the population thus far is called *global best*. According to the following equations, PSO searches for the optimal solution by updating the velocity and the position of each particle:

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

$$v_{id}(t+1) = w * vid(t) + \mathbf{c1} * r1i * (pid - xid(t)) + c2 * r2i * (pgd - xid(t)) \quad (2)$$

where t denotes the t_{th} iteration, $d \in D$ denotes the d_{th} dimension in the search space, w is inertia weight, c_1 and $c_2$are the learning factors called, respectively, cognitive parameter, social parameter, r_{1i} and r_{2i} are random values uniformly distributed in [0, 1], and finally p_{id} and p_{gd} represent the elements of *local best* and *global best* in the d_{th} dimension. The personal best position of particle *i* is calculated as

$$y_i(t+1) = \{ y_i(t) if\ f(x_i(t+1))[?][?][?] f(y_i(t)) x_i(t+1)\ if\ f(x_i(t+1)) < f(y_i(t)) \quad (3)$$

In this work, PSO is a filter with CFS (Correlation-based Feature Selection) as a fitness function. Like the majority of feature selection techniques, CFS uses a search algorithm along with a function to evaluate the worth of feature subsets. CFS measures the usefulness of each feature for predicting the class label along with the level of inter correlation among them, based on the hypothesis: Good feature subsets contain features highly correlated (predictive of) with the class, yet uncorrelated with (not predictive of)each other [12].

D. Design of the Wrapper Phase

After filtering, the next phase is the wrapper, Wrapper methods used previously usually adopt random search strategies, such as Adaptive Genetic Algorithm (AGA), Chaotic Binary Particle Swarm Optimization (CBPSO) and Clonal Selection Algorithm (CSA). The wrapper approach consists of methods choosing a minimum subset of features that satisfies an evaluation criterion.

It was proved that the wrapper approach produces the best results out of the feature selection methods [13], although this is a time-consuming method since each feature subset considered must be evaluated with the classifier algorithm. to overcome this, we perform filtering using the PSO in advance so that we can reduce the execution time.In the wrapper method, the features subset selection algorithm exists as a wrapper around the data mining algorithm and outcome evaluation.The induction algorithm is used as a black box. The feature selection algorithm conducts a search for a proper subset using the induction algorithm itself as a part of the evaluation function. GA-based wrapper methods involve a Genetic Algorithm (GA) as a search method of subset features.

GA is a random search method, effectively exploring large search spaces [14]. The basic idea of GA is to evolve a population of individuals (chromosomes), where an individual is a possible solution to a given problem. In the case of searching the appropriate subset of features, a population consists of different subsets evolved by a mutation, a crossover, and selection operations. After reaching maximum generations, algorithms returns the chromosome with the highest fitness, i.e. the subset of features with the highest accuracy.

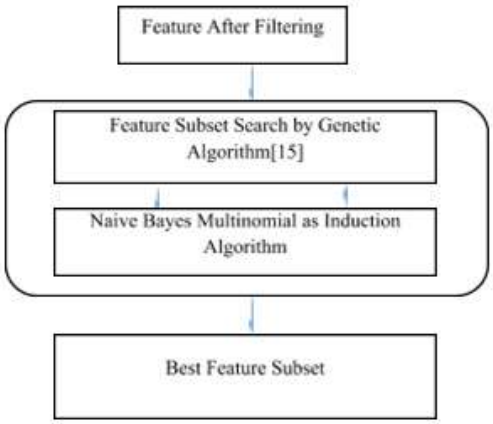

Figure 2. GA-based wrapper feature selection with Naive Bayes Multinomial as an induction algorithm evaluating feature subset.

E. Naive Bayes Multinomial

Multinomial Naive Bayes (MNB) is the version of Naive Bayes that is commonly used for text categorization problems. In the MNB classifier, each document is viewed as a collection of words and the order of words is considered irrelevant.

The probability of a class value *c* given a test document *d* is computed as

$$P(d) = \frac{P(c) \prod_{w[?][?][?]d} P(w|c)^{n_{wd}}}{P(d)} \quad (4)$$

where n_{wd} is the number of times word *w* occurs in document *d*, *P* (*w*|*c*) is the probability of observing word *w* given class *c*, *P* (*c*) is the prior probability of class *c*, and *P* (*d*) is a constant that makes the probabilities for the different classes sum to one. *P* (*c*) is estimated by the proportion of training documents pertaining to class *c* and *P* (*w*|*c*) is estimated as

$$P(c) = \frac{1 + \sum_{d[?][?][?]Dc} n_{wd}}{k + \sum_{wI} \sum_{d[?][?][?]D_c} n_{wId}} \quad (5)$$

Many text categorization problems are unbalanced. This can cause problems because of the Laplace correction used in. Consider a word *w* in a two class problem with classes *c*1 and *c*2, where *w* is completely irrelevant to the classification. This means the odds ratio for that particular word should be one, i.e. $\frac{p(w|c_1)}{p(w|c_2)}$, so that this word does not influence the class probability. Assume the word occurs with equal relative frequency 0.1 in the text of each of the two classes. Assume there are 20,000 words in the vocabulary (k = 20, 000) and the total size of the corpus is 100,000 words.

Fortunately, it turns out that there is a simple remedy: we can normalize the word counts in each class so that the total size of the classes is the same for both classes after normalization[17].

III. Experimental results

In order to evaluate the performance of the proposed method, our algorithm has been tested using 478 documents from www.shamela.ws, which are independently classified into seven categories, 70% for data Training and 30% for data testing. The number of words (features) in this dataset is 1578, in our experiments, we have successfully reduced the number of words (features) into

618 using CFS-PSO filtering algorithm and then the second phase we reduce result features filtering using wrapper Genetic Algorithm (GA) for optimizing features into 266. We have compared the accuracy our proposed method NBM-PSO-GA with NBM without filtering and NBM using Algorithm PSO for filtering with the same dataset.

3.1. The classification results of the NBM, NBM-PSO, and NBM-PSO-GA

The classification results of the NBM, NBM-PSO, and NBM-PSO-GA Classifier are shown in Tables 3.2, 3.3, and 3.4 respectively.

Tabel 3.1. Number of documents selected from Dataset

Class Name	Doc Num.
Sholat ÕáÇÉ	136
Zakat ÇáÒßÇÉ	36
Shaum ÕíÇã	42
Hajj ÇáÍÇÌ	64
mutazawwij ãÊÒæÌ	46
Baa'aand Isytaro **ÇáÈíÚ æÇáÔÑÇÁ**	109
Wakaf ÇáÃæÞÇÝ	45

As shown in Table 3.2, when the Naive Bayes Multinomial classifier is used, the Accuracy Rate, Recall Rate, F-Measure and Precision of the entire classification result Accuracy Rate are 81.8%. The table indicates that Recall is 0.82, Precession is 0.82 and F-Measure are 0.828. Moreover, the table indicates that "*Sholat*" classification Accuracy can reach 97.4 % and the Accuracy Rate of the ''*Shaum*" class is as low as 66.7% and this influences the entire classification result.

As shown in Table 3.3, when the Naive Bayes Multinomial with a features filter selection (PSO) classifier is used, the Accuracy Rate, Recall Rate, F-Measure and Precision of the entire classification result Accuracy Rate are 81.8%. The table indicates that Recall is 0.82, Precession is 0.84, and F-Measure is 0.82. Moreover, the table indicates that "*Sholat*" classification accuracy can reach 97,4% and the Accuracy Rate of the ''*Baa'aand Isytaro*" class is as low as 68,8% and this very influences the entire classification result.

Finally, the method we propose, as shown in Table 3.3. Naive Bayes Multinomial with Hybrid features selection with Filter(PSO) and Wrapper(GA) classifier is used, the Accuracy Rate, Recall Rate, F-Measure and Precision of the entire classification result Accuracy Rate are 90,2%, Recall is 0.90, Precession is 0.91 and F-Measure is 0.90. Moreover, the table indicates that "*Sholat*" classification accuracy can reach 97,z% and the Accuracy Rate of the ''*Baa'aand Isytaro*" class is as low as 71,4% and this very influences the entire classification result.

Tabel 3.2. Classification results of the Naive Bayes Multinomial

Class Name	Recall	Precision	F-Measure	Accurate
Sholat صلاة	0.97	0.93	0.95	**97.4**
Zakat الزكاة	0.82	0.69	0.75	81.8
Shaum صيام	0.67	0.75	0.71	66.7
Hajj الحاج	0.86	0.67	0.75	85.7
Mutazawwij متزوج	0.71	0.83	0.77	71.4
Baa'aand Isytaro والشراء البيع	0.75	0.84	0.79	75
Wakaf الأوقاف	0.73	0.79	0.76	73.3
Total	0.82	0.82	0.82	81.8

Tabel 3.3. Classification results of the Naive Bayes Multinomial Filtering with PSO

Class Name	Recall	Precision	F-Measure	Accurate
Sholat صلاة	0.97	0.93	0.95	**97.4**
Zakat الزكاة	0.91	0.71	0.80	90.9
Shaum صيام	0.78	0.70	0.74	77.8
Hajj الحاج	0.93	0.65	0.77	92.9

Mutazawwij م تزوج	0.71	0.71	0.71	71.4
Baa'aand Isytaro والـ شراء الـ بـ يع	0.69	0.92	0.79	68.8
Wakaf الأوقـ ـاف	0.73	0.73	0.73	73.3
Total	0.82	0.84	0.82	81.8

Tabel 3.4. Classification results of the Naive Bayes Multinomial Filtering PSO + Wrapper GA

Class Name	Recall	Precision	F-Measure	Accurate
Sholat صلاة	0.97	0.93	0.95	**97.4**
Zakat الـ زكـاة	0.91	0.83	0.87	90.9
Shaum صـ يام	0.89	0.89	0.89	88.9
Hajj الـ حاج	0.93	0.81	0.87	92.9
Mutazawwij م تزوج	0.71	0.71	0.71	71.4
Baa'aand Isytaro الـ بـ يع والـ شراء	0.88	0.96	0.91	87.5
Wakaf الأوقـ ـاف	0.87	0.93	0.90	86.7
Total	0.90	0.91	0.90	**90.2**

Conclusions of 3 The table above shows that the method that we propose to show the value of the highest accuracy is 85.90%, compared with the NBM and NBM-PSO, and the lowest for the accuracy of the 81% that NBM-PSO.

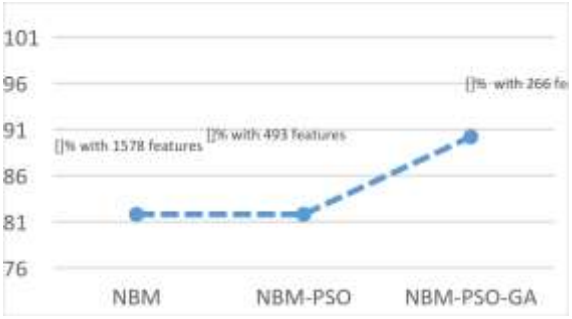

Figure 3 Accuracy with number of features

From Figure 3.4 it can be seen, features selection using Filtering PSO, successful can reduce features from 1578 into 493, with the same accuracy of the classification without any filtering, and the we propose method is the result of filtering of the PSO, we Optimizing features using GA , and the results are in addition to the number of features is reduced to 266, accuracy is also increased by 8.4%.

IV. Conclusion and future work

This paper proposed the integration of NBM-PSO-GA methods for Arabic Document Classification, our experiments revealed the important of feature selection process in building a classifier. The integration of PSO+GA has greatly reduced the features by keeping the resources to a minimum while at the same time improves the classification accuracy. The accuracy rate for our method is 85.89%. For future work, we will try to combine the two methods of classification with other selection features for optimal results.

Reference

[1]. Chakrabarti, S., 2000. Data mining for hypertext: A tutorial survey. SIGKDD explorations 1(2), ACM SIGKDD.
[2]. Salton, G., 1989. Automatic Text Procession: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading, PA.
[3]. Rifkin, R.M., 2002. Everything Old is New Again: A Fresh Look at Historical Approaches in Machine Learning, Ph.D. Dissertation, Sloan School of Management Science, September, MIT.
[4]. Salton, G., 1989. Automatic Text Procession: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading, PA.
[5]. Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, Fourteenth International Conference on Machine Learning.
[6]. John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In Proceedings of the 11th International Conference on machine learning (pp. 121–129). San Francisco: Morgan Kaufmann
[7]. A. McCallum, K. Nigam, A comparison of event models for naive Bayes text classification, in: AAAI-98 Workshop on Learning for Text Categorization, 752, 1998, pp. 41–48.

[8]. G. Forman, An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res. 3 (2003) 1289–1305.
[9]. M. Rogati, Y. Yang, High-performing variable selection for text classification, in: CIKM '02 Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, pp. 659–661.
[10]. Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: The Fourteenth International Conference on Machine Learning (ICML 97), 1997, pp. 412–420.
[11]. J .Kennedy, R.Eberhart, "Particle Swarm Optimization",In :Proc IEEE Int. Conf. On Neural Networks, Perth, pp. 1942-1948, 1995.
[12]. Matthew Settles, "An Introduction to Particle Swarm Optimization", November 2007.
[13]. X. Zhiwei and W. Xinghua, "Research for information extraction based on wrapper model algorithm," in 2010 SecondInternational Conference on Computer Research and Development, Kuala Lumpur, Malaysia, 2010, pp. 652–655
[14]. D. Goldberg, Genetic Algorithms in Search, Optimization, andMachine Learning. Addison-Wesley, Reading, MA, 1989.
[15]. Asha Gowda Karegowda, M.A.Jayaram, A.S .Manjunath."Feature Subset Selection using Cascaded GA & CFS:A Filter Approach in Supervised Learning". International Journal of Computer Applications (0975 – 8887) , Volume 23– No.2, June 2011
[16]. Hall, Mark A and Smith, Lloyd A, " Feature subset selection: a correlation based filter approach", Springer, 1997.
[17]. Eibe Frank1 and Remco R. Bouckaert ," Naive Bayes for Text Classification with Unbalanced Classes" Computer Science Department, University of Waikato,2006