ARABIC TEXT CATEGORIZATION SYSTEM
Using Ant Colony Optimization-based Feature Selection

Abdelwadood Moh'd A. Mesleh and Ghassan Kanaan
Faculty of Information Systems & Technology, Arab Academy for Banking and Financial Sciences, Amman, Jordan

Keywords: Arabic Text Classification, Feature Selection, Ant Colony Optimization, Arabic Language, SVMs.

Abstract: Feature subset selection (FSS) is an important step for effective text classification (TC) systems. This paper describes a novel FSS method based on Ant Colony Optimization (ACO) and Chi-square statistic. The proposed method adapted Chi-square statistic as heuristic information and the effectiveness of Support Vector Machines (SVMs) text classifier as a guidance to better selecting features for *selective* categories. Compared to six classical FSS methods, our proposed ACO-based FSS algorithm achieved better TC effectiveness. Evaluation used an in-house Arabic TC corpus. The experimental results are presented in term of macro-averaging F_1 measure.

1 INTRODUCTION

It is known that the volume of Arabic information available on the Internet is increasing. This growth motivates researchers to better classifying Arabic articles. TC (Manning & Schütze, 1999) is the task to classify texts to one of a pre-specified set of categories based on their contents.

Arabic TC process compromises three main components (Mesleh, 2007): data pre-processing, text classifier construction, and document categorization. *Data pre-processing* makes the text documents compact and applicable to train the text classifier. *Text classifier construction* implements the function of learning from a training dataset. After evaluating the effectiveness of the text classifier, the TC system can implement the function of Arabic *document classification*. When given an enough number of labeled examples (training dataset), we can build a TC model to predict the category of new documents. Those examples include a huge number of features, and some of the features do not reveal significant document-category characteristics. This is why FSS techniques are essential to decrease the size of training dataset, to speed up training process and to improve the text classifier's effectiveness.

In this paper, much attention is paid to *pre-processing* and in particular to the FSS process.

The rest of this paper is organized as follows. In section 2, an overview of FSS methods is presented; section 3 describes our proposed Ant Colony Optimization-based FSS method (ACO-based FSS). Experimental results and conclusions are discussed in sections 4 and 5 respectively.

2 FSS OVERVIEW

FSS is a process that chooses a subset of features from an original feature set according to some criterion, the FSS basic steps are (Liu & Yu, 2005):

Feature Generation. In this step, a number of candidate subsets of features are generated by some search process.

Feature Evaluation. In this step, the candidate feature subsets are evaluated to measure their goodness. Evaluation is divided into filter and wrapper methods. In filter methods, features are selected by a filtering process that is based on scores which were assigned by a specific weighting method. On the other hand, in wrapper methods, feature selection is based on the accuracy of some given classifier.

Stopping Criteria. In this step, the FSS process stops if a predefined criterion is met.

In TC tasks, many FSS approaches (Yang & Pedersen, 1997; Forman, 2003) are often used such as Document Frequency thresholding (DF), Chi-square statistic (CHI), Term Strength (TS),

Moh'd A. Mesleh A. and Kanaan G. (2008).
ARABIC TEXT CATEGORIZATION SYSTEM - Using Ant Colony Optimization-based Feature Selection.
In *Proceedings of the Third International Conference on Software and Data Technologies - PL/DPS/KE*, pages 384-387
DOI: 10.5220/0001892803840387

Information Gain (IG), Mutual Information (MI), Odds Ratio (OR), NGL coefficient and GSS score.

The valuable FSS studies (Yang & Pedersen, 1997; Forman, 2003) investigated FSS methods for English TC tasks. However, Syiam, Fayed and Habib (Syiam, Fayed & Habib, 2006) evaluated the effectiveness of many FSS methods (Chi-square, DF, IG, OR, GSS score, and NGL coefficient) for Arabic TC tasks with Rocchio and *k*NN classifiers. They concluded that a hybrid approach of DF and IG is a preferable FSS method for Arabic TC task.

In a recent FSS study, Mesleh (Mesleh, 2007) has conducted an empirical comparison of these FSS methods evaluated on an Arabic dataset with SVMs classifier. Mesleh concluded that Chi-square works best with SVMs classifier for Arabic TC tasks.

Theoretically FSS has been shown to be an NP-hard problem (Blum, & Rivest, 1992), as a result, automatic feature space construction and FSS from a large set has become an active research area. On the other hand, optimization algorithms (such as genetic algorithm (Goldberg, 1989)) have become applicable to FSS processes.

In this work, ACO algorithm is proposed to enhance (optimize) the Chi-square based FSS process; the following may justify the selection of ACO algorithm for Arabic FSS in TC task:

- Comparing with other evolutionary-based algorithms, ACO algorithm (Elbeltagi, Hegazy & Grierson, 2005) performs better in term of processing time. Where processing time is very important when dealing with the huge number of features in TC Arabic dataset.
- Compared to English, Arabic language (Yahya, 1989) is more sparsed, which means that English words are repeated more often than Arabic words for the same text length. Sparseness yields less weight for Arabic terms (features) than English features. The difference of weight among Arabic word features is less and this makes it more difficult to differentiate between different Arabic words (this may negatively affect Arabic text classifier's effectiveness).

ACO algorithm, which imitates foraging behavior of real life ants (Dorigo, Maniezzo & Colorni, 1996) was first proposed to solve traveling salesman problem. However, it has been recently proposed to solve many other problems such as FSS.

3 PROPOSED ACO-BASED FSS

ACO algorithm was used (Al-Ani, 2005) in the FSS processes for speech segment and texture classification problems. Similarly, ACO algorithm was used (Jensen, & Shen, 2003) in an entropy-based modification of the original rough set-based approach for FSS problems. ACO algorithm was used (Schreyer, & Raidl, 2002) to label point features, a pre-processing step to reduce the search space. And a hybrid method (Sivagaminathan, & Ramakrishnan, 2007) of ACO and Neural Networks was used to select features.

The main difference between these FSS approaches is in the calculation of the used heuristic values. Heuristic values help the algorithm reach an optimal solution. Accordingly, we have tailored ACO algorithm to fit the FSS process for Arabic TC tasks. This new proposed FSS method adapted Chi-square statistic as heuristic information and the effectiveness of SVMs text classifier as a guidance to better selecting features for selective text categories in our Arabic TC system.

The main steps of our proposed ACO-Based FSS method are as follows:

Initialization Step. Initially, Ant colony algorithm parameters are initialized:

- Define the amount of pheromone change for each feature $\Delta\tau_i = 0$. Where i is a feature index, $i \in [0, N]$, and N is the total number of features in the feature space.
- Define pheromone level associated with each feature ($\tau_i = 1$).
- Stopping criterion: define the maximum number of iterations (NIs = 30).
- Define the desired macro-averaging F_1 measure ($BF_1 = 88.11$).
- Define the number of Solutions (number of ants) (NAs = 30).
- Define the number of features in each candidate subset of features (NFs).
- Define the Number of Top Best Solutions (TBS=10).
- Local Selection Criterion: for all the features in the original feature set, Chi-square statistic scores are pre-calculated.

Step 2 – Generation Ants for Initial Iteration. FOR each ant (solution) ($ant_i : i = 1 : NAs$), randomly select NFs features.

Step 3 – Evaluation Solutions. FOR each solution ($ant_i : i = 1 : NAs$), run the classifier (SVMs text

classifier) to evaluate the goodness of solution i, Evaluation is based on macro-averaging F_1 measure

Step 4 – Stopping Criterion. IF a predefined stopping criterion is met THEN stop the Ant Colony Optimization-Based FSS process.
ELSE:

(1) Pheromone Update. Update the pheromone levels associated with features in the TBS solutions, pheromone update is defined by:

$$\tau_i = \begin{cases} \rho\tau_i + \Delta\tau_i + w.\Delta\tau_i & f_i \in EBS_j \\ \rho\tau_i + \Delta\tau_i & \text{otherwise} \end{cases}$$

Where:

- $\Delta\tau_i$ is defined by:

$$\Delta\tau_i = \begin{cases} \dfrac{\max\limits_{g=1:TBS}(F_{1g}) - F_{1j}}{\max\limits_{h=1:TBS}(\max\limits_{g=1:TBS}(F_{1g}) - F_{1h})} & f_i \in S_j \\ 0 & \text{therwise} \end{cases}$$

- ρ is a coefficient such that $(1-\rho)$ represents the *evaporation* of pheromone level. Elitist Best Solution (EBS_j) is any solution S_j among the TBS solutions that outperformed BF_1. w is the performance effectiveness of solution S_j. And f_i is a feature indexed by i.

(2) Probabilistic Feature Selection. Select new features for the NAs ants for the next iteration. Selection is defined by the following Chi-square based Feature Selection Probability (CHIFSP):

$$CHIFSP_i^{S_j} = \begin{cases} \dfrac{[\tau_i]^\alpha \cdot [CHI_i^{S_j}]^\beta}{\sum\limits_{g\ is\ allowed} [\tau_g]^\alpha \cdot [CHI_g^{S_j}]^\beta} & f_i \notin S_j \\ 0 & \text{otherwise} \end{cases}$$

Where:

- CHI_i^{Sj} is the local importance of feature f_i given the solution S_j.
- α and β are used to control the effects of Chi-square statistic and the pheromone level.
- Go to evaluation Step 3.

4 EXPERIMENTAL RESULTS

In this work, we have used an in-house collected corpus (see Mesleh, 2007). It consists of 1445 documents of different lengths belonging to nine categories. We followed (Mesleh, 2007) in processing the Arabic dataset: Each article in the Arabic dataset is processed to remove digits and punctuation marks. Normalize some Arabic letters such as "ء" (hamza) in all its forms to "ا" (alef). All the non Arabic texts were filtered. Arabic function words (such as "آخر", "أبدا", "أحد" etc.) were removed. Arabic documents were represented by vector space model. Lastly, all terms with term frequency less than some threshold were filtered (threshold is set to *three* for positive features and set to *six* for negative features in training documents).

To implement SVMs text classifier (Mesleh, 2007), we used an SVMs package, TinySVM (downloaded from http://chasen.org/~taku/), the soft-margin parameter C is set to 1.0. And in order to fairly compare our ACO-Based FSS with other FSS methods, six FSS methods (IG, CHI, NGL, GSS, OR and MI) were implemented. For the ACO-Based FSS method, α and β are set to 1.

TC effectiveness (Baeza-Yates and Ribeiro-Neto, 1999) is measured in terms of Precision, Recall and F_1 Measure. Denote the precision, recall and F_1 measures for a class C_i by P_i, R_i and F_i, respectively. We have:

$$\mathrm{P_i} = \frac{TP_i}{TP_i + FP_i},\ \mathrm{R_i} = \frac{TP_i}{TP_i + FN_i},\ \mathrm{F_i} = \frac{2\mathrm{P_i}\mathrm{R_i}}{\mathrm{R_i} + \mathrm{P_i}}$$

Where TP_i, FP_i, FN_i, and TN_i are defined in Table 1.

Table 1: The Contingency Table for Category c_i.

Category c_i		Expert Judgment	
		YES	NO
Classifier Judgment	YES	TP_i	FP_i
	NO	FN_i	TN_i

To evaluate the average performance over many categories, the macro-averaging F_1 (F_1^M) is used and defined as follows:

$$\mathrm{F_1^M} = 2[\sum_{i=1}^{|C|}\mathrm{R_i}\sum_{i=1}^{|C|}\mathrm{P_i}] / N[\sum_{i=1}^{|C|}\mathrm{R_i} + \sum_{i=1}^{|C|}\mathrm{P_i}]$$

To evaluate the effectiveness of our proposed ACO-based FSS method, we conducted three groups of TC experiments. For each group and for each text category, we have randomly specified one third of the articles and used them for testing while the remaining articles used for training the Arabic classifier. And for each FSS method (ACO-based FSS, Chi-square, GSS, NGL, IG, OR, and MI), we have conducted three experiments to select 180, 160, and 140 features respectively. Then we conducted an additional experiment without any FSS method (the result of this experiment is referred to as *original* classifier). In this work, *ONLY* one category's

features are selected by ACO-based FSS method, i.e. SVMs F_1^M results were achieved by *only* optimizing *one* text category (the smallest category). We noted that optimizing any category will enhance the classifier's effectiveness.

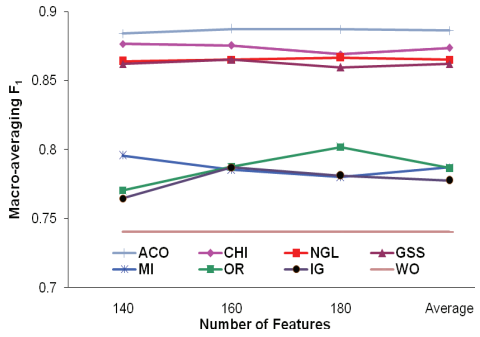

Figure 1: SVMs F_1^M values for SVMs with the seven FSS methods at different subset of features.

Figure 1 shows F_1^M results for SVMs text classifier with the seven FSS methods at different sizes of feature subsets. It is obvious that our ACO-based FSS method outperformed the *original* classifier (where all the 78699 features are used for training the SVMs text classifier) and outperformed the other six FSS methods. Best Chi-square F_1^M result was 88.11, and after optimizing the feature selection of the smallest category, F_1^M result became 88.743.

5 CONCLUSIONS

Our proposed ACO-based FSS method adapted Chi-square statistic as heuristic information and the effectiveness of SVMs as a guidance to better selecting features in Arabic TC tasks. In this work, the proposed FSS method was selectively applied to a single text category (*Computer* category is the smallest category). Compared to six classical FSS methods, it achieved better TC effectiveness results. Optimizing features for all categories, tuning the ACO-based FSS parameters and studying their effects, and comparing our proposed method with other ACO algorithm flavors are left as future work.

REFERENCES

Manning, C., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Liu, H., Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 4, 491-502.

Yang, Y., Pedersen, J., 1997. A Comparative Study on Feature Selection in Text Categorization. In J. D. H. Fisher, editor, The 14th International Conference on Machine Learning (ICML'97), Morgan Kaufmann, 412-420.

Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research, vol. 3, 1289-1305.

Syiam, M., Fayed, Z., Habib, M., 2006. An Intelligent System for Arabic Text Categorization. International Journal of Intellegent Computing & Information Ssciences, vol.6, no.1, 1-19.

Mesleh, A., 2007. Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study, to appear in the proceedings of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 07), December 3-12, Springer-Verlag.

Blum, A., & Rivest, R.., 1992. Training a 3-Node Neural Network is NP-Complete. Neural Networks, vol. 5, no. 1, 117-127.

Goldberg, D., 1989. Genetic Algorithms in search, optimization, and machine learning, Addison-Wesley.

Dorigo, M., Maniezzo, V., Colorni A., 1996. The ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B, vol. 26, no. 1, 29--41.

Elbeltagi, E., Hegazy, T., Grierson, D., 2005. Comparison among five evolutionary-based optimization algorithms, Advanced Engineering Informatics, vol. 19, no. 1, 43-53.

Yahya, A., 1989. On the complexity of the initial stages of Arabic text processing, First Great Lakes Computer Science Conference; Kalamazoo, Michigan, USA.

Al-Ani, A., 2005. Feature Subset Selection Using Ant Colony Optimization, International Journal of Computational Intelligence. vol. 2, no. 1, 53-58.

Jensen, R., Shen, Q., 2003. Finding rough set reducts with ant colony optimization. In Proceedings of the 2003 UK workshop on computational intelligence, 15-22.

Schreyer, M., Raidl, G., 2002. Letting ants labeling point features. In Proceedings of the 2002 IEEE congress on evolutionary computation at the IEEE world congress on computational intelligence, 1564-1569.

Sivagaminathan, R.K., Ramakrishnan, S., 2007. A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert Systems with Applications, vol. 33, 49-60.

Baeza-Yates, R., Rieiro-Neto, B., (1999). Modern Information Retrieval. Addison-Wesley & ACM Press.