World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:12, 2010

A Proposed Hybrid Approach for Feature Selection in Text Document Categorization

M. F. Zaiyadi and B. Baharudin

***Abstract*—**Text document categorization involves large amount of data or features. The high dimensionality of features is a troublesome and can affect the performance of the classification. Therefore, feature selection is strongly considered as one of the crucial part in text document categorization. Selecting the best features to represent documents can reduce the dimensionality of feature space hence increase the performance. There were many approaches has been implemented by various researchers to overcome this problem. This paper proposed a novel hybrid approach for feature selection in text document categorization based on Ant Colony Optimization (ACO) and Information Gain (IG). We also presented state-of-the-art algorithms by several other researchers.

***Keywords*—**Ant colony optimization, feature selection, information gain, text categorization, text representation.

I. Introduction

FROM manual storage, text document is now available in digital format. With the massive growth of the World Wide Web, users can now access the documents they want through the Internet. The development of the documents is growing rapidly which contributes to large amount of data available in the Internet. This situation brings difficulty to users as they have to choose the best document that suits their search. Then efforts taken by domain experts to manually categorize the documents. But as the number of documents increased, it was quite a waste of time and energy. Nowadays there are many computer applications that can make document selection rather easier and faster by performing text categorization. Text document categorization refers to an automatic process of classifying text document into one of several fixed classes or categories.

In text categorization system development, documents are represented as feature vectors before it can be made suitable for the classification. Then these feature vectors are divided into two set, which is the training set and test set. Features in the training set will be used to train the learning algorithm and build the classifier. The classifier is then used to classify the incoming text into pre-determined classes. For evaluation purpose, the classifier is applied on the test set and the result is evaluated to see the performance of the classification. Classifier performance will be evaluated using macro-precision, macro-recall and macro-f1 and hence evaluation of the feature selection algorithm effectiveness**.**

Typically, text is represented as bag-of-words. One of the main difficulties that have to be dealt with before performing the classification is the high dimensionality of the feature space. The feature space consists of tens or hundreds of thousands of unique terms extracted from input documents which also include irrelative and noisy data. These characteristics can hurt the performance of the classifier and consequently affecting the result. In order to prevent this situation, the extracted features have to be filtered prior to classification phase as to select the most relevant ones that can best represent the documents. This is done by removing non-informative terms and constructing new set of features using automatic feature selection methods.

There were many algorithms introduced and applied in the machine learning field especially for feature selection. In this paper, we present several state-or-the-art applications of feature selection algorithm from the data mining field particularly in text document categorization. Then we propose one potential novel hybrid algorithm for feature selection which attempt to optimize the advantages of ant colony optimization and information gain, two of the most popular algorithms for feature selection.

Ant colony optimization (ACO) is an algorithm that can be used to solve combinatorial optimization problems. It is a metaheuristic in which a colony of artificial ants cooperates in finding good solutions to difficult discrete optimization problems [9]. It mimics the behavior of real ants' colony in their effort of searching for the shortest path from their food source to their nest. With the advantages that ACO has, many researchers have been carrying out researches involving ACO in various fields of applications. Meanwhile, information gain (IG) is term-goodness criterion which has been used frequently in machine learning problems. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document.

The remainder of the paper is organized as follows: Section 2 discusses on the background of the study which includes some applications of ant colony optimization and information gain; Section 3 describes the proposed approach and Section 4 concludes this paper.

II. Research Background

A. Ant Colony Optimization

Many methods have been implemented for feature selection

M. F. Zaiyadi is with the Computer and Information Science Department at the Universiti Teknologi Petronas, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia (e-mail: fairuz_zaiyadi@yahoo.com).

B. Baharudin is with the Computer and Information Science Department at the Universiti Teknologi Petronas, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia (email: baharbh@petronas.com.my).

and mostly involved statistical approaches. However with the advancement of knowledge and technology, many other methods from different field have also been applied for feature selection. Among others, population-based optimization algorithms such as genetic algorithm (GA) and ant colony optimization (ACO) have attracted much interest.

Ant colony optimization is a novel nature-inspired meta-heuristic for the solution of hard combinatorial optimization (CO) problems which was first introduced by M. Dorigo and his colleges in early 1990s. The idea was inspired by the foraging behavior or real ants' colonies [16] and was originally used as an artificial system to produce near-optimal solution for the classical traveling salesman problem [12] before it was successfully applied for many other difficult problems such as the quadratic assignment problem, network routing, scheduling, etc [10].

In the year of 2005, Zhand and HU applied ACO with Mutual Information (MI) in a hybrid approach for feature selection in the forecaster [11]. Then Zhou et al. [12] presented a hybrid method to select optimum feature subset in Equipment Fault Diagnosis also using ACO and MI. In [13], Ming proposed a hybrid feature selection method which combines rough set theory and ACO. Meanwhile in the year of 2009, Basiri and Nemati [14] successfully applied ACO as a hybrid method with Genetic Algorithm (GA) to get an excellent feature set for text categorization problem. They proved that the proposed algorithm was easily implemented with low computational cost.

Additional to that, [15] proved that ACO is an appropriate algorithm when dealing with large number of features in text categorization. They made a comparative study between five optimization algorithms for discrete optimization problems and found that particle swarm optimization generally outstanding in terms of success rate and solution quality. However, ACO performs better in term of processing time and that is important for high number of features.

B. Information Gain

The main concept of a feature selection is to apply an evaluation function towards each feature in the initial feature collection extracted from a text document to measure the degree of relevancy of a feature to a document. There are numbers of algorithms developed and applied for automatic selection of features. The most popular and reliable one is information gain as proved by several researchers. Apart from that, there are other traditional methods which have been widely used for feature selection.

Document Frequency (DF), Chi-Square (CHI), Information Gain (IG), Mutual Information (MI), Term Strength (TS) and Odds Ratio are among the commonly used evaluation functions for feature selection in various applications [1]. Tasci and Gungor in [2] also claimed that Information Gain (IG), Chi-Square (CHI) and Document Frequency (DF) were among popular feature selection metrics and further proved that IG gave the best results in most of their experiments. Yang *et al*. [3] experimented five score functions on Reuters 21578 collection and found that DF, CHI and IG were the most effective among all.

Other than that, Gabrilovich *et al*. [4] made an experiment of text categorization with many redundant features using several different feature selection algorithms and finally found out that IG, CHI and BNS were the best performers. As in other application, Sheen and Rajesh [5] considered three different feature selection approaches for network intrusion detection and they showed that IG and CHI were in better performance. So we can see here that in most experiments, IG is considered as one the best solution for feature selection

In terms of hybrid application involving IG, there were many researches carried out combining IG with other techniques to select the best features before performing classification. In 2004, Li *et al*. [6] introduced hybrid method for feature selection in topic-specific text filtering using several feature selection methods which includes CHI, MI and IG, combined with rough set theory. Meanwhile, Wang *et al*. [7] adopted hybrid method to feature selection in Chinese Text Sentiment Classification based on category distinguishing ability of words and information gain. Then a year later, Yang *et al*. [8] implemented information gain and chaotic genetic algorithm as a hybrid technique to select relevant genes in micro-array data classification problem.

However, information gain also has some drawbacks and one of them is that it does not perform well and less appropriate in a highly redundant set of features [18]. Therefore we propose to overcome this problem by performing a hybrid method which combines ACO and IG.

III. Proposed Approach

In this paper, we proposed a hybrid approach for feature selection which combines ACO and IG to reduce the dimensionality of the feature space. We are trying to make full use of both methods in selecting features by combining them to develop a more promising algorithm which would then result in better classification of text document.

A. Ant Colony Optimization

The ACO algorithm is based on a computational paradigm inspired by real ant colonies and the way they function. When searching for food, ants communicate each other using pheromone, passing information regarding the shortest path to the food source. A single moving ant lays some pheromone on the ground, thus making a path by a trail of this substance. When other ant moves at random and encounter previously laid trail, it can detect and decide with high probability to follow it by laying its own pheromone and consequently reinforces the trail. The more the ants follow a trail, the more attractive that trail becomes to be followed. Thus the process is characterized by a positive feedback loop where the probability of choosing a path increases with the number of ants choosing the same path [17].

However, feature selection problem has to be reformulated by means of making it ACO-suitable task. ACO requires a problem to be represented as a graph where features are

represented as nodes with the edge between them denoting the choice of the next feature. Traversal of an ant through the graph is considered as the search for the optimal feature subset where a minimum number of nodes are visited that satisfies the traversal stopping criterion. Fig. 1 illustrates this setup.

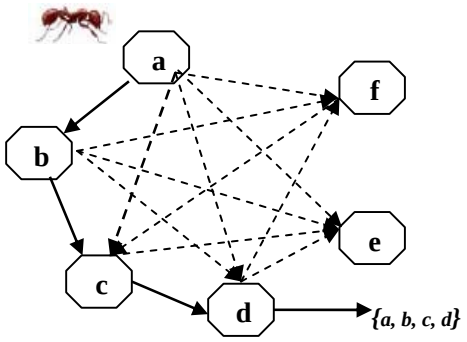


Fig. 1 ACO representation for FS problem

Referring to fig. 1, nodes are fully connected to allow any feature to be selected next. The ant is initially at node *a*, and has a choice of choosing the next feature to be added to its path. Based on the transition rule, it chooses *b* as the next feature, then *c* and then *d*. Upon arrival at node *d*, the current subset {*a, b, c, d*} is determined to satisfy the traversal stopping criterion. The ant terminates its traversal and outputs this feature subset as a candidate for data reduction [17].

In constructing solution, prior to traversal of the graph, all ants have to be put in empty memory condition which is done by setting all arrays to false. Then each ant has to be assigned an initial node or feature at random. They construct a complete tour by applying a choice rule at each construction steps based on heuristics information and pheromone trails. As long as the stopping criterion is not satisfied, the ants will traverse through the nodes looking for the best feature for the subset.

An artificial ant follows a rule called probabilistic transition rule which determines the probability of an ant *k* choosing feature *i* to be added to it's solution at time t. The probabilistic transition rule is build of two parameters which is heuristic desirability and pheromone levels as follows:

$$p_i^k(t) = \begin{cases} \dfrac{[\tau_i(t)]^{\alpha} \cdot [\eta_i]^{\beta}}{\sum_{j \in J^k} [\tau_j(t)]^{\alpha} \cdot [\eta_j]^{\beta}} & \text{if } i \in J^k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where J^k is the set of ant k's unvisited features, η_i is the heuristic desirability of choosing feature *i* to be part of the partial solution, $\tau_i(t)$ is the pheromone value laid at feature *i*, while α and β are two parameters that determine the relative importance of the pheromone value and heuristic information.

After all ants have completed their solutions, the pheromone produced during the tour will have to be managed. There are two procedures involves in pheromone management, namely pheromone evaporation and pheromone deposit. Both procedures are comprised in one main procedure, pheromone update. The function of pheromone evaporation is to make sure that the ants are not traversing on the same path, constructing the same solution. Then all ants can update the pheromone level on the features they have visited and the best ant with the best solution will get the chance to deposit more pheromone than others.

B. Information Gain

As mentioned previously, information gain is among the frequently employed evaluation function for term or feature selection in machine learning field. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The information gain of a term *t* is defined as:

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) \\ & + P_r(t) \sum_{i=1}^{m} P_r(c_i \mid t) \log P_r(c_i \mid t) \\ & + P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i \mid \bar{t}) \log P_r(c_i \mid \bar{t}) \end{aligned} \quad (2)$$

where $P_r(c_i)$ is the probability of a document to fall under the class label $P_r(t)$ is the probability of a term *t* to appear in a document, $P_r(c_i \mid t)$ is the probability of a document to fall under class label c_i given that term *t* appears in the document and $P_r(c_i \mid \bar{t})$ is the probability of a document to fall under class label c_i given that term *t* does not appear in the document.

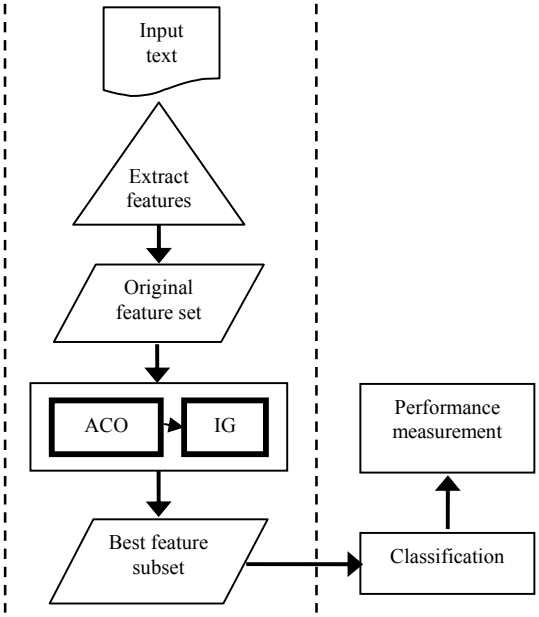
C. Hybrid Feature Selection: ACO-IG

In previous sections, we have briefly explained on the advantages of IG as one of the best and most popular algorithm for feature selection especially in text categorization. We also mentioned that one of the drawbacks of IG is that it is less appropriate when operated with highly redundant features. Meanwhile, the ability of ACO as a feature selection algorithm has also been discussed in various field of applications. It was claimed that ACO is appropriate for high dimensional data and was shown to be an effective tool in finding good solutions [24]. Thus we are looking into possibilities of combining these two algorithms together, forming a hybrid feature selection by means of optimizing the advantages of both method.

Process of the proposed approach is illustrated in fig. 2. As shown in the figure, the original feature set extracted from input documents will be first filtered by ACO to reduce the redundancy among features. Then the feature subset produced by ACO will be further evaluated by IG to select the best features to represent the document. The best feature subset is then used for the classification.

Fig. 2 Block diagram of feature selection with ACO-IG

The overall process of ACO-IG can be seen in fig. 3.



According to fig.3, the process begins with the generation of ants which are then put randomly on the graph that represents the candidate features. The number of ants to place on the graph may be set equal to the number of features and each ant starts their traversal from a different feature. Prior to the path construction, we will do the initialization of the number of ants, pheromone level for each candidate feature, termination condition and other parameters for the algorithm. Then from the initial node, the ants with empty memory will traverse through the graph probabilistically and construct a complete tour based on pheromone trails and heuristic information.

The pheromone level at each feature is the result of the pheromone deposited by an ant during its traversal which is proportional to the quality of the solution and regarded as prior knowledge for other ant. Meanwhile, heuristic desirability of an ant traversing between features could be interpreted by any subset evaluation function such as entropy-based measure [19] or rough set dependency measure [20]. In this proposed algorithm, heuristic desirability for feature selection is determined using IG as in (3) below:

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{[\tau_i(t)]^{\alpha} \cdot [IG_i]^{\beta}}{\sum_{j \in J^k} [\tau_j(t)]^{\alpha} \cdot [IG_j]^{\beta}} & \text{if } i \in J^k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

From (3), we can see that the heuristic desirability of candidate feature i is calculated by IG_i which combined with pheromone trails level τ_i to determine whether to select feature i to be part of the partial solution or not. The result of the traversal is gathered and the subset will be applied with IG to further select the best features that can best represent the text. Then the resulting subsets are evaluated. When the ants satisfy the stopping criterion, the process halts with the best feature subset to be used for the classification. Else, the process will continue with the next iterations by updating the level of pheromone on each visited features and new ants are generated for the new selection of features.

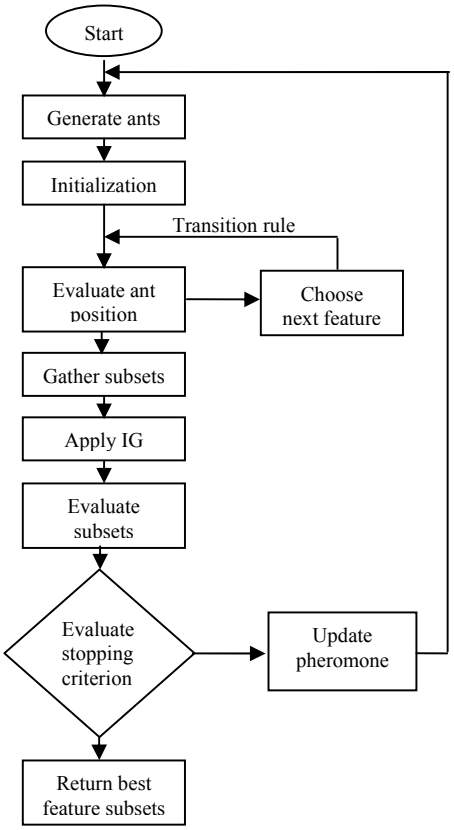


Fig. 3 Flow diagram of ACO-IG feature selection

IV. CONCLUSION

IG has been proved to be one of the best algorithms for feature selection in text categorization. Meanwhile, population-based methods have also triggered interests among researchers in machine learning and one of the most promising algorithms is ACO. Many experiments have been carried out showing that ACO performs well in solving problems especially in feature selection. Thus we propose a hybrid method for feature selection in text categorization adopting ACO and IG, looking into the possibilities of increasing the performance of IG by optimizing the strength of ACO.

For now, we have identified the potential combinations of both methods for feature selection which illustrated through the flow diagram presented in this paper. Further works will include the development of a detail algorithm for the proposed feature selection approach, realization of the algorithm and categorization of documents in order to evaluate the effectiveness of the proposed approach.

REFERENCES

[1] F. Sebastiani, "Machine learning automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1 – 47, March 2002.
[2] A. Tasci and T. Gungor, "An evaluation of existing and new feature selection metrics in text categorization", *International Symposium on Computer and Information Science*, pp. 1-6, Oct. 2008.

[3] Y. Yang and J. O. Pedersen, "A Comparative study on feature selection in text categorization", *Proceeding of 14th International Conference on Machine Learning,* San Francisco, 1997, pp. 412-420.
[4] E. Gabrilovich and S. Markovitch, "Text Categorization with many redundant features: using aggressive feature selection to make SVM competetive with C4.5", *Proceeding of 21st International Conference on Machine Learning*, Canada, 2004.
[5] Sheen and Rajesh, "Network intrusion detection using feature selection and decision tree classifier", *IEEE Region 10 Conference*, Hyderabad, pp. 1-4, Nov. 2008.
[6] Q. Li, J.H. Li, G.S. Li, and S.H. Li, "A rough set-based hybrid feature selection method for topic-specific text filtering", *Proceedings of the Third International Conf. on Machine Learning and Cybernetics,* Shanghai, August 2004, pp. 1464-1468.
[7] S. Wang, Y. Wei, and D. Li, "A hybrid method of feature selection for Chinese text sentiment classification", *Fourth International Conf. on Fuzzy Systems and Knowledge Discovery,* 2007.
[8] C.S. Yang, L.Y. Chuang, J.C. Li, and C.H. Yang, "Information gain with chaotic genetic algorithm for gene selection and classification problem", IEEE International Conference on Systems, Man and Cybernetics, pp. 1128-1133, Oct. 2008.
[9] M., Dorigo and T. Stutzle, *Ant Colony Optimization*, MIT press, 2004, pp.25-26.
[10] H.R. Kanan, K. Faez and M. Hosseinzadeh, "Face recognition system using ant colony optimization-based selected features", *IEEE Symposium on Computational Intelligence in Security and Defense Applications*, pp. 57-62, Apr. 2007.
[11] C.K. Zhang and H. Hu, "Feature selection using the hybrid of ant colony optimization and mutual information for the forecaster", *Proceedings of the Fourth International Conf. on Machine Learning and Cybernetic,* Guangzhou, August 2005, pp. 1728-1732.
[12] J. Zhou, R. Ng, and X. Li, "Ant colony optimization and mutual information hybrid algorithms for feature subset selection in equipment fault diagnosis", *10th International Conf. on Control, Automation, Robotics and Vision*, Hanoi, Vietnam, December 2008.
[13] M. He, "Feature selection based on ant colony optimization and rough set theory", *International Symposium on Computer Science and Computational Technology*, pp. 247-250. Dec. 2008.
[14] M.E. Basiri and S. Nemati, "A novel hybrid ACO-GA algorithm for text feature selection", *IEEE Congress on Evolutionary Computation*, pp. 2561-2568, 2009.
[15] E. Elbeltagi, T. Hegazy and D. Grierson, "Comparison among five evolutionary-based optimization algorithms", *Advanced Engineering Informatics*, vol. 19, no. 1, pp. 43-53, 2005.
[16] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey", *Theoretical Computer Science*, pp. 243-278, 2005.
[17] M.H. Aghdam, N.G. Aghaee and M.E. Basiri, "Application of ant colony optimization for feature selection in text categorization", *IEEE Congress on Evolutionary Computation*, pp. 2867-2873, June 2008.
[18] C. Lee and G.G. Lee, "MMR-based feature selection for text categorization", *Proceedings of the Annual Conf. of Human Language Technology conference / North American chapter of the Association for Computational Linguistic,* May 2004.
[19] R. Jensen, "Combining rough and fuzzy sets for feature selection", *Ph.D. Dissertation*, School of Information, Edinburgh Univ., 2005.
[20] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishing, Dordrecht, 1991.
[21] A.M. Mesleh and G. Kanaan, "Support vector machine text classification system: Using ant colony optimization based feature subset selection", *Int. Conf. on Computer Engineering and Systems*, pp. 143-148, Nov. 2008.
[22] M. Sadeghzadeh and M. Teshnehlab, "Correlation based feature selection using ant colony optimization", *World Academy of Science, Engineering and Technology 64*, pp. 497-502, 2010.
[23] A. Al-Ani, "Ant colony optimization for feature subset selection", *World Academy of Science, Engineering and Technology 4*, pp. 35-38, 2005.
[24] M. Deriche, "Feature selection using ant colony optimization", *International Multi-Conference on Systems, Signals and Devices*, pp. 1-4, March 2009.
[25] L. Wen, Q. Yin, and P. Guo, "Ant colony optimization algorithm for feature selection and classification of multispectral remote sensing image", *IEEE Int. Geosciences and Remote Sensing Symposium*, pp. 923-926, July 2008.
[26] W. Xiong and C. Wang, "A hybrid improved and colony optimization and random forest feature selection method for microarray data", *Fifth International Joint Conference on INC, IMS and IDC*, pp. 559-563, 2009.