Genetic Algorithm based Feature Selection in High Dimensional Text Dataset Classification

FERHAT ÖZGÜR ÇATAK
TÜBİTAK - BİLGEM
Cyber Security Institute
Kocaeli Gebze
TURKEY
ozgur.catak@tubitak.gov.tr

Abstract: Vector space model based bag-of-words language model is commonly used to represent documents in a corpus. But this representation model needs a high dimensional input feature space that has irrelevant and redundant features to represent all corpus files. Non-Redundant feature reduction of input space improves the generalization property of a classifier. In this study, we developed a new objective function based on models F_1 score and feature subset size based. In this paper, we present work on genetic algorithm for feature selection in order to reduce modeling complexity and training time of classification algorithms used in text classification task. We used genetic algorithm based meta-heuristic optimization algorithm to improve the F_1 score of classifier hypothesis. Firstly; (i) we've developed a new objective function to maximize; (ii) then we choose candidate features for classification algorithm; and (iii) finally support vector machine (SVM), maximum entropy (MaxEnt) and stochastic gradient descent (SGD) classification algorithms are used to find classification models of public available datasets.

Key–Words: Feature selection, support vector machines, logistic regression, stochastic gradient descent, document classification

1 Introduction

This research is motivated by a complex model representation problem of language model for text classification task. Our research focus mainly on the reducing model complexity of classifier functions using reducing the size of input feature vector. We applied the proposed model to the publicly available text classification datasets that contain thousands of features.

Text classification is a special case of classification task in machine learning field [1]. In recent years, text classification has gained considerable attention. It is one of the most attractive topics in sentiment analysis [2], information retrieval [3]. Main difference with other classification task of other domains is that text classification data set contains large number of features. Most of the text classification research uses bag of words model that contains unique word or phrases to convert that occur in documents in a corpus to vector space based language model. But the main problem with the bag of words technique is that it generates hundreds or thousands features in input space which is not efficient way of vectorizing features. Although the language model contains high feature space, most of the generated features of this technique are irrelevant, redundant or noisy [4]. Filtering non-redundant and irrelevant features from text data set is a primary problem for high dimensional input spaces in this field. Features with high discriminative power build robust classification models and also they reduce the computational complexity of training algorithms [5].

Another advantage of feature selection is improvement of model accuracy and the stability. Feature selection on classifier models has great impact on predictive results with high dimensional input space. Main advantage of feature selection in the machine learning is to reduce the complexity of classification hypothesis. And also, feature selection provides avoiding over fitting of classifier function [6], reducing memory consumption of learning algorithm [7] and removes the irrelevant features (terms) from text data set [8].

Feature selection methods are basically divided into two different approaches: filter and wrapper methods [9]. Filter methods based feature selection algorithms use an evaluation function to perform feature selection [10]. Each feature is ranked according to a metric. Feature rank metric can be fisher score, t–test, or information gain. This method is independent with the classifier algorithms. Filter methods are very

efficient and fast to compute. But this method can select feature subset, which has redundant features. The second option for feature selection is wrapper methods [11]. Wrapper methods based feature selection algorithms use subset of feature set and accuracy of the classifier function that is trained with these subset of feature set based training data. This method is dependent with the classifier algorithm.

In recent years, heuristic algorithms have been used widely in feature selection for large-scale data sets. Kabir et.al. [12] developed ant colony optimization based feature selection method for neural network algorithm. Their approach combines both advantages of wrapper and filter based methods. They compare their results with other existing feature selection algorithms. Chen et.al [13] developed also ant colony optimization with rough set theory based feature selection method. Their approaches uses mutual information based feature significance. They used 9 different UCI datasets which of feature size are from 7 to 70. Unler et.al [14] developed a different approach with particle swarm optimization algorithm based feature selection. Their feature selection method uses relevance and dependence of the features included in the feature subset. They are also used public dataset for results. Bae et.al [15] developed a new method to overcome premature converge of objective function. Their particle swarm based approach uses intelligent swarms for heuristic search.

This research proposed a new algorithm that considers both the number of features in feature subset and F_1 score of the classifier function that is generated with this feature subset. F_1 score is the most used model selection method in information retrieval domain. The overall contribution of the study can be listed as follows:

1. Using feature selection, the input matrix which is quite high for the memory is reduced, and, input matrix complexity is reduced in this manner.

2. F_1 based model selection method is used.

3. Iteration size of our algorithm is quite low.

The rest of the paper is organized as follows. Section 2.1 provides a brief review of SVM, MaxEnt and SGD classification algorithms. Section 2.2 provides genetic algorithm based heuristic optimization. Section 3 presents the detail of proposed feature selection method. Section 4 gives the experimental results with public datasets. Last section discusses the experimental results and the future works of the proposed model.

2 Preliminaries

In this section, we will give brief information about classification methods in Section 2.1 and genetic algorithm in Section 2.2.

2.1 Classification Methods

In our experiments, we've used three different classification algorithm; support vector machine, maximum entropy and stochastic gradient descent. In this section, we will give some brief information about these machine learning algorithms.

2.1.1 Support Vector Machine

Support vector machine (SVM) classification algorithm is widely used supervised learning method in machine learning field. SVM is based on statistical learning theory and tries to maximize the generalization property of classifier model that is generated by algorithm. SVM classification algorithm uses a set of training instances and predicts new instances with two possible class label $-1, 1$. As shown in Figure 1, the hyperplane is defined by $w^T x + b = 0$, where $w \in R^n$ is a orthogonal to the hyperplane and $b \in R^n$ is the constant. Giving some training data D, a set of point of the form.

$$D = \{(\vec{x}_i, y_i) \,|\vec{x}_i \in R^m, y_i \in \{-1, +1\}\}_{i=1}^n \quad (1)$$

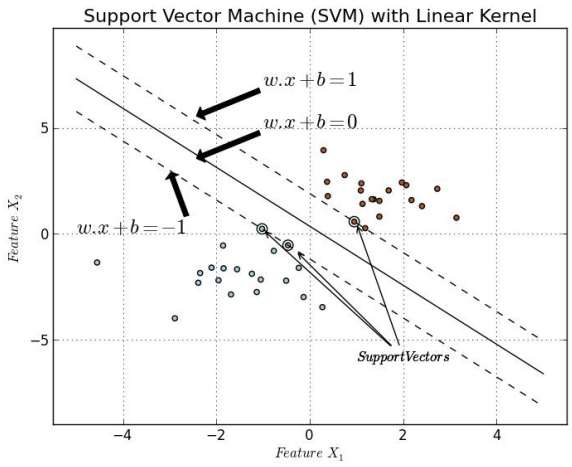
where x_i is a m-dimensional real vector, y_i is the class of input vector x_i either -1 or $+1$. SVM aims to search a hyper plane that maximizes the margin between the two classes of samples in D with the smallest empirical risk [16]. For the generalization property of SVM, two parallel hyperplanes are defined such that $w^T x + b = 1$ and $w^T x + b = -1$. One can simplify these two functions into new one.

$$y_i \left(\vec{w}^T \vec{x} + b\right) \geq 1 \quad (2)$$

SVM aims to maximize distance between these two hyperplanes. One can calculate the distance between these two hyperplanes with $\frac{1}{||\vec{w}||}$. The training of SVM for the non-separable case is solved using quadratic optimization problem that shown in Equation 3.

$$\begin{aligned} &minimize : P\left(\vec{w}, b, \xi\right) = \frac{1}{2}||\vec{w}|| + C\sum_{i=1}^{n} \xi_i \\ &subject\,to : y\left(\vec{w}.\phi\left(\vec{x}\right) + b\right) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0 \end{aligned} \quad (3)$$

Figure 1: SVM classification algorithm separating hyper plane illustration.



for $i = 1, ..., m$, where ξ_i are slack variables and C is the cost variable of each slack. C is a control parameter for the margin maximization and empirical risk minimization.

2.1.2 Maximum Entropy

Maximum entropy (MaxEnt) is another linear classification model based on empirical data. The MaxEnt is used as a means of estimating probability distributions from data X. Conditional distribution of training dataset X is used as constraints [17]. Maximum-likelihood distributions function in exponential form:

$$p\left(y|\vec{x}\right) = \frac{1}{Z\left(\vec{x}\right)} exp\left(\sum_i \lambda_{i,c} F_{i,c}\left(\vec{x}, y\right)\right) \quad (4)$$

where $Z(\vec{x})$ is normalization function. $F_{i,c}$ is a feature/class function for feature f_i and outcome y defined in form [18]:

$$F_{i,c}\left(d, c'\right) = \begin{cases} 1, & n_i\left(x\right) > 0 \text{ and } c' = c \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

2.1.3 Stochastic Gradient Descent

Stochastic gradient descent algorithm is another classification method. let $\vec{x}$ be an arbitrary instance of training dataset. y is the scalar output of instance $\vec{x}$. Our aim is the minimizing of loss function $L(\hat{y}, y)$ that measures the cost of predicting $\hat{y}$ with known outcome y. Classification task search function spaces F of $f_w(\vec{w})$ parametrized by a weight vector $\vec{w}$. Gradient descent algorithm uses empirical risk to find out parameter vector $\vec{w}$. Empirical risk can be computed with an approximation:

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} l\left(\hat{y}, y\right) \quad (6)$$

Gradient descent is used to minimize the empirical risk $E_n(f_w)$ at each iteration updating the weight vector $\vec{w}$ with a learning rate λ.

$$\vec{w}_{t+1} = \vec{w}_t - \lambda \frac{1}{n} \sum_{i=1}^{n} \Delta_w l\left(f\left(\vec{w}\right), y\right) \quad (7)$$

Stochastic gradient descent algorithm is simplified version of Equation 7.

$$\vec{w}_{t+1} = \vec{w}_t - \lambda \Delta_w l\left(f\left(\vec{w}_t\right), y_t\right) \quad (8)$$

Instead of using empirical risk, R_{emp}, SGD uses random picked instances of training set X, to estimate the gradient.

2.2 Genetic Algorithm

Genetic algorithm (GA) is an evolutionary algorithm that mimics the natural selection, crossover and mutation process. GA was first developed by Holland in 1975. GA is a stochastic optimization method, which is based on metaheuristic search procedures. GA starts with a matrix of population of solution. Each row of this matrix shows the individuals that generated randomly. Each individual shows a solution of an objective function. In GA, every solution is encoded with genes that is called individual. Using a objective function, fitness of individuals are computed according to an objective function. Population is improved with combination of genetic information from different members of population. This process is called as crossover. Another population improvement method is mutation. Some individuals of population are mutated according to the mutation rate of population. Pseudo code of GA is show in Algorithm 1.

3 Proposed Model

In this study, at the first step, we have generated uniform random population which their genes have normal distribution with $\mu = 0.3$ and $\sigma = 0.15$. Each generation has 0.3 mutations and 0.3 crossover probability. Elitism is used to retain several highest individuals to the next generation directly. *Step 1 Initial population generation:* Our model start with creating random initial population. Each chromosome contains

Algorithm 1 Genetic Algorithm

procedure GENETICALGORITHM(P)
  $t \leftarrow 0$
  $InitPopulationP\,(t)$ ▷ Initialize Population randomly
  $F(t) = ComputeFitness\,(P\,(t))$
  while $not\,terminated$ **do**
    $t \leftarrow t + 1$
    $P(t) \leftarrow crossover(P(t-1))$
    $P(t) \leftarrow mutate(P(t))$
    $F(t) \leftarrow ComputeFitness(P(t))$
  end while
  return $best\,p$ ▷ Return the best individual
end procedure

number of feature genes and each gene has real number. General representation of chromosome is shown in Equation 9.

$$C = \{\vec{f_i} | \vec{f_i} \in [0,1]\}_{i=1}^{m} \tag{9}$$

Figure 2 shows the chromosome representation of input feature set of training data set. We propose a new

Figure 2: Chromosome representation of feature set.

threshold value to select feature. The threshold function drastically converges to 0.5 with iteration steps. Our threshold function is:

$$v = exp(-2t) * rand() + 0.5 \tag{10}$$

where t is iteration number and $rand()$ is uniform random distribution function with range $[0, 1]$.

Step 2 Classifier model generation: Each chromosome in population represents selected features in training and test set. Support vector machine, maximum entropy and stochastic gradient descent classification algorithms are used to show the accuracy performance of the proposed model.

Step 3 Objective function: The objective function to be maximized is the sum of feature ratio and F_1 score of the best chromosome. F_1 measure is harmonic mean of precision and recall score of classifier. Precision is the probability that retrieved instances classified by classifier function are relevant and recall is the probability that relevant instances are retrieved by classifier function. Precision and recall scores are defined in Equation 11 and 12 respectively.

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

The $F-score$ is a harmonic mean of the precision and recall scores specially used in machine learning and information retrieval. F_1 score are defined in Equation 13.

$$F_1 = \frac{2.P.R}{P + R} \tag{13}$$

Our contributions are the objective function used in genetic algorithm based optimization for feature selection. We combine F_1 score with feature ratio as objective function. The objective function is defined in Equation 14.

$$U = \frac{\text{Num of Feature}}{\text{Num of Feature at iteration t}} + F_{1,t} \tag{14}$$

Pseudo code of the method is show in Algorithm 2.

Algorithm 2 Proposed genetic algorithm based feature selection method

$C, T, f_{best} \leftarrow \emptyset$ ▷ Initialize
$P_0 \leftarrow$ random gauss distribution with $\mu = 0.3\,and\,\sigma = 0.15$
Convert each individual gene to binary discrete such that $\begin{cases} 0, & p_i < 0.5 \\ 1, & \text{otherwise} \end{cases}$
while $t \leq T$ **do**
  $t \leftarrow t + 1$
  $GeneticAlgorithm(P_t)$
  $if\,argmax_t\,(P_t) \geq then\,f_{best} \leftarrow argmax_t(P_t)$
end while
return $best\,p$ ▷ Return the best individual

4 Results

In this section, we present the results of three different public text datasets to compare the proposed feature selection method F_1 score with original dataset. In this study we used three different public benchmark datasets to verify its model effectivity and efficiency.

We experiment on three public data sets which are summarized in Table 1, including Farm-Ads [19],

Table 1: Description of the testing data sets used in the experiments.

Dataset	Train	Test	Class	#Att.
Farm-Ads	4,000	143	2	54,877
News20	18,000	1,996	2	1,355,191
RCV1	20,242	677,399	2	47,236

News20-Binary [20] and RCV1 [21]. All experiments are repeated 5 times and the results are averaged.

In this study, support vector machine, maximum entropy and stochastic gradient descent models were constructed.

The support vector machine parameters used to find out classifier model summarized by the following:

- *Kernel:* linear
- *C*=0.01
- *Loss Function*=l_2 regularization
- *Dual mode*=True
- *Tolerance*=0.0001

The maximum entropy parameters used to find out classifier model summarized by the following:

- C=1
- Loss Function=l_2 regularization
- Dual mode=True

The stochastic gradient descent parameters used to find out classifier model summarized by the following:

- *C*=1
- *Loss Function*=hinge
- *Regularization term*=0.0001
- *Learning rate* = 0.01

Genetic algorithm based feature selection method's parameters are summarized by the following

- *Initial:* normal random generated $\in [0, 1]$
- *Population*= 30
- *Number of generation*= 1000
- *Crossover rate*= 0.9
- *Mutation rate*= 0.01
- *Elites*= 1

All datasets are separated that first 90% of the instances being used for training, and the next 10% for testing. Proposed method is developed using Python language with *scikit-learn* and *inspyred* library. Results are shown in Table 2, 3.

We showed the feature size, F_1 score and accuracy of the all datasets in Table 1-6. Accuracy changes of the proposed model are show in Figure 3. As shown in figures, there is an inverse correlation between initial population size and rate of convergence of classifier model accuracy. Our experiments for each datasets show that final accuracy level of the classifier models is same for some initial population size. For instance, in the farm-ads and RCV1 datasets feature selection experiment, final performance of the classifier model accuracy is same for 50, 150, and 200 population size. After exceeding the population size 200, classifier models accuracy smoothly becomes to decrease. Population step size of News20 dataset is different from other datasets. We choose 10 step size and initial population starts from 10 and last size is 50. Although step size and initial population size is different from others, Figure 3 shows us that classifier models accuracy is same for a known size and then its accuracy becomes to decrease.

5 Conclusion

In this work, a novel objective function is developed for feature selection task in machine learning area. Main contribution of this work is that our method especially suitable for information retrieval and text classification area to remove noisy and irrelevant features from input space of dataset while improve the performance of text classification. Our method tries to find a feature subset as small as possible while classifier hypothesis has high F_1 score. As seen in tables and figures, all training datasets are uniformly converges to the optimal classifier accuracy. Our observations show that the population size of genetic algorithm affects directly performance of the global classifier function.

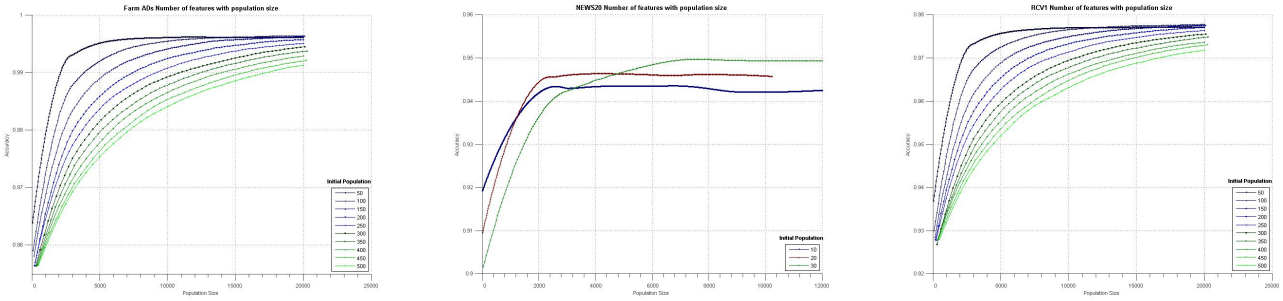
In the future, our method will be applied to more datasets for testing performance. We plan to find a relation between population sizes, iteration size of optimal converge and F_1 score of classifier model.

Table 2: SVM simulation results of selected public datasets

	All Features					Selected Features				
		Train DS		Test DS			Train DS		Test DS	
Dataset	Feat. Size	F_1	Acc.	F_1	Acc.	Feat. Size	F_1	Acc.	F_1	Acc.
Farm-Ads	54877	0.992	0.991	0.993	0.993	21627	0.984	0.983	0.976	0.974
News20-Binary	1355191	0.877	0.878	0.847	0.856	531653	0.886	0.887	0.857	0.870
RCV1	47236	0.954	0.953	0.954	0.952	18566	0.946	0.944	0.937	0.932

Table 3: Maximum entropy simulation results of selected public datasets.

	All Features					Selected Features				
		Train DS		Test DS			Train DS		Test DS	
Dataset	Feat. Size	F_1	Acc.	F_1	Acc.	Feat. Size	F_1	Acc.	F_1	Acc.
Farm-Ads	54877	0.999	0.999	0.998	0.998	21587	0.998	0.998	0.994	0.994
News20-Binary	1355191	0.967	0.967	0.960	0.960	531907	0.946	0.946	0.928	0.93
RCV1	47236	0.979	0.978	0.978	0.978	18600	0.964	0.963	0.959	0.957



(a) Farm-Ads (b) news20. (c) RCV1

Figure 3: Feature reduction over population size.

References:

[1] Colace, F., De Santo, M., Greco, L., and Napoletano, P., Text classification using a few labeled examples, *Computers in Human Behavior* 30, 2014, pp. 689–697.

[2] Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M., Building emotional dictionary for sentiment analysis of online news, *World Wide Web* 17, 2014, pp. 723-742.

[3] El-Bakry, Hazem M., and Nikos Mastorakis., Fast information retrieval from web pages. *Proceedings of the 7th WSEAS international conference on Computational intelligence, man-machine systems and cybernetics*, 2008, pp. 229–247.

[4] Buck, Christian, Kenneth Heafield, and Bas van Ooyen., N-gram counts and language models from the common crawl., *Proceedings of the Language Resources and Evaluation Conference*, 2014.

[5] Y. Guo, G. Zhao, and M. PietikäInen, Discriminative features for texture description,*Pattern Recogn.*, 45, 2012, pp. 3834-3843.

[6] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images, *Neuroimage*60, 2012, pp 59-70.

[7] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling, Feature selection using linear

Table 4: Stochastic gradient descent simulation results of selected public datasets.

	All Features					Selected Features				
		Train DS		Test DS			Train DS		Test DS	
Dataset	Feat. Size	F_1	Acc.	F_1	Acc.	Feat. Size	F_1	Acc.	F_1	Acc.
Farm-Ads	54877	0.982	0.981	0.977	0.978	21517	0.992	0.991	0.985	0.984
News20-Binary	1355191	0.991	0.990	0.999	0.999	531963	0.971	0.971	0.987	0.987
RCV1	47236	0.987	0.987	0.997	0.997	18661	0.972	0.971	0.983	0.983

classifier weights: Interaction with classification models, *in Proceedings of SIGIR 04*, 2004, pp. 234-241.

[8] L. Yu and H. Liu, Efficient feature selection via analysis of relevance and redundancy, *J. Mach. Learn. Res.* 5, 2004, pp. 1205-1224.

[9] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C. S. Cruz, Quadratic programming feature selection, *J. Mach. Learn. Res.* 11, 2010, pp. 1491-1516.

[10] Chandrashekar, Girish, and Ferat Sahin., A survey on feature selection methods., *Computers & Electrical Engineering* 40, 2014, pp. 16–28.

[11] Song, Qinbao, Jingjie Ni, and Guangtao Wang., A fast clustering-based feature subset selection algorithm for high-dimensional data., *Knowledge and Data Engineering, IEEE Transactions on* 25, 2013, pp. 1–14.

[12] M. M. Kabir, M. Shahjahan, and K. Murase, A new hybrid ant colony optimization algorithm for feature selection, *Expert Systems with Applications* 39, 2012, pp. 3747–3763.

[13] Y. Chen, D. Miao, and R. Wang, A rough set approach to feature selection based on ant colony optimization, *Pattern Recognition Letters* 31, 2010, pp. 226-233.

[14] A. Unler and A. Murat, A discrete particle swarm optimization method for feature selection in binary classification problems, *European Journal of Operational Research* 206, 2010, pp. 528–539.

[15] C. Bae, W.-C. Yeh, Y. Y. Chung, and S.-L. Liu, Feature selection with intelligent dynamic swarm and rough set, *Expert Syst. Appl.* 37, 2010, pp. 7026-7032.

[16] Vapnik, Vladimir., *The nature of statistical learning theory*, 1995

[17] Nigam, Kamal, John Lafferty, and Andrew McCallum.,Using maximum entropy for text classification, *IJCAI-99 workshop on machine learning for information filtering.* 1, 1999.

[18] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan.,Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* 10, 2002, pp. 79–86.

[19] Chris Mesterharm, Michael J. Pazzani, Active Learning using On-line Algorithms, *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011,pp. 850–858.

[20] Ken Lang., Newsweeder: Learning to filter netnews, *In Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339

[21] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li., RCV1: A new benchmark collection for text categorization research, *Journal of Machine Learning Research* 5, 1995, pp. 331–339