Twenty Newsgroups Text Classification

Final Report

Lu Zhao[1] and Yan Liang[1]

[1]Department of Applied Mathematics and Statistics, Stony Brook University

***Abstract*—Feature selection is usually used to deal with the high dimensionality of feature space in text classification. In our project, we consider three criteria of feature selection: document frequency, mutual information and χ^2 statistic. The general selection process is done in a greedy manner, at each step, the new feature with highest score (e.g. χ^2 statistic) condition on a subset of features already selected will be selected. In addition, we propose to simulated annealing algorithm to select new feature at each step. We will explore the relationship between results of greedy and simulated annealing algorithms and to what extent can we reduce the feature space without significantly damage the classification accuracy. Our experimentation results shows that simulated annealing algorithm is more effective that greedy algorithm and with very little amount of features (about 0.1%) we can achieve about 60% of test accuracy using entire feature space.**

***Index Terms*—Test classification, Feature selection**

I. Introduction

Text classification is the problem of assigning a text with a label from the predefined label set and there are extensive literature applying different machine learning techniques. Naive Bayes method is considered in [2], EM algorithm in [3] and SVM is studied in [1]. Other models including neural networks, decision tree and regression models are also applicable.

In text classification, one main challenge is the high dimensionality of the feature space. For example, in the bag of words model, the feature space consists of every word that is in the text. As a results, the dimension of the feature space can be hundreds of thousands. In our project, the original feature space contains 129,797 words for ten thousands of documents. This high dimensionality will usually place a hardship on machine learning algorithms. In Bayes models, if the conditional independence assumptions do not holds, then the number of parameters which will increase exponentially will make the algorithm computationally unworkable. To deal with this situation, feature selection is usually considered to reduce the dimension of the feature space in order to increase the efficiency of the classification algorithm and at the same time minimize the loss of classification accuracy. One characteristic of feature selection is that it is feature oblivious, the feature is selected not based on a priori information but the information inscribed in the data.

Feature selection contains two general methods, one is scoring each feature and then extracting a subset of feature space; the other one is latent feature extraction, e.g. principal component analysis. In our project, we focus on the first feature selection method. The widely used statistics are eliminating rare words, document frequency, estimated mutual information, χ^2 statistic and information gain [4], [5]. The feature selection is usually done in a greedy way, i.e. at each step, pick the feature that achieves highest score (mutual information, χ^2 statistic, etc.) conditioning on the features previously selected. Since this is not maximizing (or minimizing which depends on the statistics in use) a global objective function, although each time a local optimal feature is selected, the overall performance of the combination of the features may not be desirable. Based on this fact, we propose to use simulated annealing algorithm which may select sub-optimal feature at each step. Some experiments in other fields show a promising results of simulated annealing algorithm [6]. Other method we are considering is doing pair-wise comparison of all subsets of features. The challenge is that the complexity is huge for pair-wise comparison due to the number of combinations of features which is exponential in number of features as well as the non-trivality of the calculation of the score for each combination which will dramatically increase running time. So, we will apply simulated annealing algorithm, which is a heuristic algorithm for large scale NP-complete problem. Even the optimality of the solution given by simulated annealing algorithm is not guaranteed, it has the potential to avoid the local optimal solution.

In this project, we are trying to answer the following two questions

1) What is the relation between selection statistics and selection scheme? For several statistics like mutual information or χ^2 statistic, will them perform differently under greedy or simulated annealing selection scheme? If so, which is better.
2) If a certain amount decreasing of classification accuracy is allowed, then how much features can we reduce such that classification accuracy is still acceptable compared to classification using original feature space.

After feature selection, we are using three classifiers to evaluate the validity of the feature we selected. They are SVM, naive bayes and nearest neighbor algorithms. In the following parts of project final report, we present the method we are using in section II: document frequency, mutual information, χ^2 statistic and simulated annealing algorithm. In section III, performance of our proposed method is presented. In section IV, the conclusions are reached.

II. Methodology

In our report, we evaluate three selection statistics: document frequency, mutual information and χ^2 statistic.

Document frequency is the most naive statistic. The document frequency of a feature (word) is just the count of texts in which this word appeared at least once. In addition, if a subset of feature space has already been selected, then the document frequency of a new feature (word) given this subset of features is the count of texts in which every feature in the subset as well as the new word appear at least once. The underling assumption of these selection statistic is that if a word appears less often, it is less informative and less influential. What's more, if a word is very rare, then it may be noise or outliers.

A. Mutual Information

Mutual information is the expectation of the point-wise mutual information which measures the mutual dependence of two random variables defined as follow: Let X denote the feature (word) and Y denote the labels

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}),$$

where $p(x,y)$ is the joint probability distribution of X and Y and $p(x), p(y)$ are the probability distribution function of X and Y, respectively. The density is estimated using MLEs:

$$\hat{P}(X=k) = \hat{p}(k) = \frac{n_{X,k}}{N}, \;\; k \geq 0,$$

where $n_{X,k}$ is the number of texts that word X appears k times and N is the total number of texts in training data

$$\hat{P}(Y=j) = \hat{p}(j) = \frac{n_j}{N}, \;\; j \in \{1,2,\ldots,24\},$$

where n_j is the number of texts with label j.

After a subset Z of features is selected, the mutual information between a new feature X and the labels Y is calculated using conditional mutual information that is conditioned on Z, i.e.

$$\begin{aligned} &I(X;Y|Z) = E_Z[I(X;Y)|Z] \\ &= \sum_{x,y,z} p_{X,Y,Z}(x,y,z)\log(\frac{p_Z(z)p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)}), \end{aligned}$$

where the density is estimated by MLEs.

Mutual information also has the following properties: i.

1) $I(X;Y) = 0$ if and only if X and Y are independent;
2) $I(X;Y)$ is non-negative;
3) and $I(X;Y) = I(Y;X)$.

The underling assumption for mutual information is that the more the label is dependent on a feature, then that feature is more informative and influential in text classification. As a consequence, in feature selection, it is desirable to select the feature that has larger mutual information evaluated with respect to the label under the condition of the feature already selected.

B. χ^2 statistic

The test of independence consists of two outcomes of variables, one is feature and one is label. The null hypothesis is that the occurrence of two outcomes are statistically independent. Assume feature X takes I values and label Y takes J values, and O_i and O_j denote the frequency of $X = i$ and $Y = j$, respectively, and $O_{i,j}$ denote the frequency of $\{X = i, Y = j\}$. Then, the estimated expected frequency is

$$E_{i,j} = \frac{O_i O_j}{N},$$

where N is the total number of sample size.

The test statistic is

$$\chi^2 = \sum_{i\in I}\sum_{j\in J}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

and the degree of freedom is $(I-1)(J-1)$.

After a subset Z of features is selected, the conditional independence is tested using the summation of all χ^2 test statistics of X and Y across all levels of Z. Assume Z takes K values. Then the degree of freedom of the new test statistic is $\sum_{k\in K}(I_k - 1)(J_k - 1)$ where I_k and J_k are the number of possible values of X and Y given that $Z = k$.

The reason of use of χ^2 statistic for test of independence is similar to that of mutual information. If a feature and the label are more dependent, then it is more informative for text classification.

C. Feature Selection

Given a scoring method mentioned above, the feature selection is done by iterating

1) choose single highest scoring feature X_k;
2) rescore all features, conditioned on the set of features already selected.

D. Simulated Annealing

The simulated annealing algorithm is applied when finding the highest scoring feature. Instead, starting from the initial temperature, each time a random feature is selected, if the score is improving, then the new feature is accepted. On the other hand, if score is not improving, then we accept the feature with certain probability decided by current temperature. Then the temperature cools down at given cooling rate and repeat the above process until the final temperature is reached.

The key configurations of simulated annealing algorithm are

1) Score: Since simulated annealing deals with minimization problem, then for document frequency and mutual information the score is negated, for χ^2 statistic, the score is measured by p-value which is a minimization problem;
2) Initial temperature (t_{ini}): the initial temperature is decided such that at the beginning of the process, any new feature could be selected. In our implementation, $t_{ini} = 450$;
3) Final temperature (t_{final}): when the temperature is around final temperature, there should be more iterations

than at the beginning of the cooling procedure. In our implementation, $t_{final} = 0.1$;

4) Number of iterations (n): In our implementation, $n = 1000$;
5) Cooling rate (r): cooling rate is calculated by

$$r = N \times \sqrt{\frac{t_{final}}{t_{ini}}}.$$

III. Main Results

A. Data Set

We extract the features from twenty newsgroups using bag of words model. There are 129,797 features. All the texts are partitioned into training data set which contains 11,314 texts with the frequency of each category shown in table 1 and test data set which contains 7,532 texts with frequency of each category shown in table 2. Texts from each category are approximately evenly distributed.

TABLE 1
Text frequency of each category in training data set.

label	count	percent (%)
0	480	4.24
1	584	5.16
2	591	5.22
3	590	5.21
4	578	5.11
5	593	5.24
6	585	5.17
7	594	5.25
8	598	5.29
9	597	5.28
10	600	5.30
11	595	5.26
12	591	5.22
13	594	5.25
14	593	5.24
15	599	5.29
16	546	4.83
17	564	4.98
18	465	4.11
19	377	3.33
total	11314	100

B. General Classifiers with Entire Feature Space

In our experimentation, we implement three popular classifiers: nearest neighbor, naive bayes and SVM (with linear kernel). From each classifier, we use ten-fold cross validation to determine the parameter. The test accuracy for each classifier is shown in table 3. For k-nearest neighbor classifier, number of neighbor used is $k = 2$; For Naive Bayes, the regularization constant is $\alpha = 0.01$; For SVM, the penalty constant is $C = 1$.

From the test accuracy, we can see that SVM and Naive Bayes are better than nearest neighbor classifier.

C. Experiment Results

In the experimentation, we use document frequency (DF) to score the feature. Then 100 features are selected through greedy feature selection as described in section II-C and

TABLE 2
Text frequency of each category in test data set.

label	count	percent (%)
0	319	4.24
1	389	5.16
2	394	5.23
3	392	5.20
4	385	5.11
5	395	5.24
6	390	5.18
7	396	5.26
8	398	5.28
9	397	5.27
10	399	5.30
11	396	5.26
12	393	5.22
13	396	5.26
14	394	5.23
15	398	5.28
16	364	4.83
17	376	4.99
18	310	4.12
19	251	3.33
total	7532	100

TABLE 3
Text Classification Test Accuracy

	k-nearest neighbor	Naive Bayes	SVM
Accuracy	0.41	0.76	0.77

simulated annealing feature selection as described in section II-D. For simulated annealing feature selection, since it's a randomized algorithm, then we repeat 10 times and 10 different sets of 100 features are generated. After feature selection, we do the classification using k-nearest neighbor, Naive Bayes and SVM. Then we compare the test accuracy between two methods, one is the common method using greedy feature selection and the other is the method we proposed using simulated annealing. From the following results, it can be seen that simulated annealing feature selection outperforms greedy selection.

1) Greedy Feature Selection with Document Frequency: We use greedy feature selection method to select 100 features which achieves highest conditional document frequency score. Then three classifiers (k-nearest neighbor, Naive Bayes and SVM) are used to test the feature selected. The test accuracy is shown in figure 1 where the classification is done using the first 1, 5, 10, 15, 20, 30 ,40, 80 and 100 features. The test accuracy is strictly increasing when number of features is increasing for Naive Bayes and SVM, for k-nearest neighbor the test accuracy may decrease as the number of features increase and it always achieves the worst test accuracy among three classifier. In addition, SVM always outperforms Naive Bayes in the case tested.

2) Simulated Annealing Feature Selection with Document Frequency: We use simulated annealing feature selection method to select 100 features which achieves sub-optimal conditional document frequency score. We run ten replications of simulated annealing feature selection since it's a randomized algorithm. Then three classifiers (k-nearest neighbor, Naive

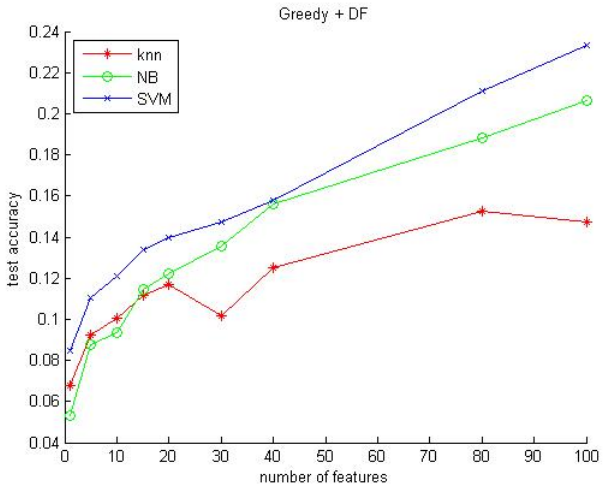

Fig. 1. Test accuracy of three classifiers, k-nearest neighbor, Naive Bayes and SVM. Feature used in classification is selected by greedy feature selection method and document frequency (DF) statistic. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The test accuracy is increasing as the number of features used increases in most cases. In addition, SVM always has the best accuracy and Naive Bayes follows, k-nearest neighbor is the worst.

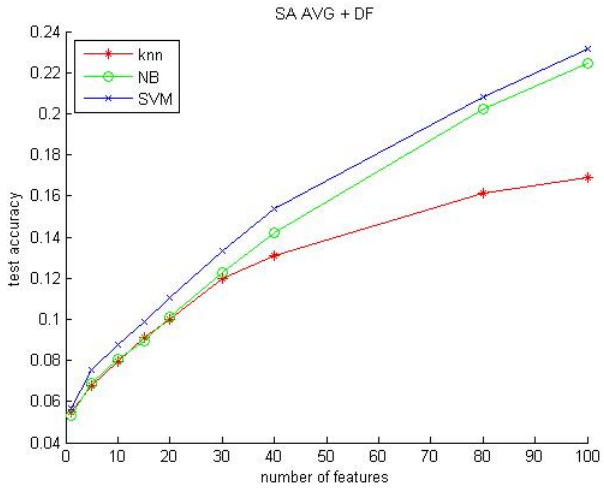

Fig. 2. The average of ten replication test accuracy of three classifiers, k-nearest neighbor, Naive Bayes and SVM. Feature used in classification is selected by simulated annealing feature selection method and document frequency (DF) statistic. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The average test accuracy is increasing as the number of features used increases for three classifier. In addition, SVM always has the best accuracy and Naive Bayes follows, k-nearest neighbor is the worst.

Bayes and SVM) are used to test the feature selected. Two kinds of test accuracy are calculated, one is the average test accuracy of ten sets of features obtained from each replication and the other one is the maximum test accuracy. The average and maximum test accuracy is shown in figures 2 and 3, respectively, where the classification is done using the first 1, 5, 10, 15, 20, 30 ,40, 80 and 100 features.

The average test accuracy keeps increasing when number of features is increasing for all three classifiers. In addition, SVM always outperforms Naive Bayes which beats k-nearest neighbors in the case tested. In addition, when more features are used, SVM and Naive Bayes tend to have similar performance, while the increment of test accuracy of k-nearest neighbor is relatively small compared to the other two classifiers.

The maximum test accuracy also keeps increasing when number of features is increasing for all three classifiers as in the average test accuracy case. In addition, SVM always outperforms Naive Bayes which beats k-nearest neighbors in most case tested. In addition, when more features are used, SVM and Naive Bayes tend to have similar performance, while the increment of test accuracy of k-nearest neighbor is relatively small compared to the other two classifiers. When number of features used is small, the fluctuate of test accuracy between Naive Bayes and k-nearest neighbor may due to the randomness of the simulated annealing algorithm and on average Naive Bayes is better, as shown in the previous plot.

In the following section we will compare the results of the method we proposed (simulated annealing feature selection) and the greedy feature selection method.

3) Comparison between Greedy and Simulated Annealing with k-Nearest Neighbor: In this section we compare the feature selected by greedy and simulated annealing algorithms using k-nearest neighbor classifier. The test accuracy using feature selected by greedy algorithm, average test accuracy

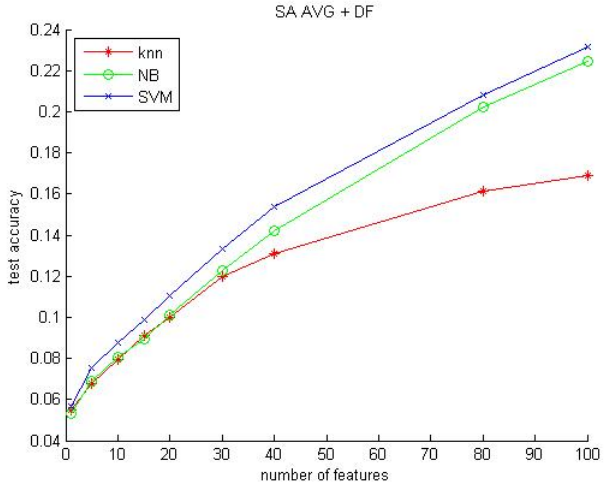

Fig. 3. The maximum of ten replication test accuracy of three classifiers, k-nearest neighbor, Naive Bayes and SVM. Feature used in classification is selected by simulated annealing feature selection method and document frequency (DF) statistic. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The maximum test accuracy is increasing as the number of features used increases for three classifier. In addition, SVM always has the best accuracy and Naive Bayes outperforms k-nearest neighbor with more features are used while losses when number of features is small.

and maximum test accuracy of ten replications of simulated annealing algorithm are shown in figure 4. From the plot, we can see that the feature selected by simulated annealing algorithm is better than greedy feature selection on average test accuracy as well as maximum test accuracy when the number of features is large.

4) Comparison between Greedy and Simulated Annealing with Naive Bayes: In this section we compare the feature selected by greedy and simulated annealing algorithms using

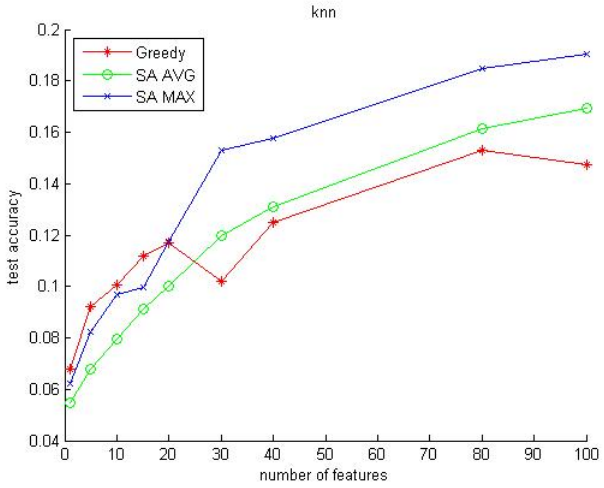

Fig. 4. Comparison of test accuracy between greedy and simulated annealing feature selection methods with document frequency (DF) statistic using k-nearest neighbor classifier. 'Greedy' denotes the test accuracy using the feature selected by greedy feature selection method. 'SA AVG' and 'SA MAX' denote the average and maximum test accuracy, respectively, of ten replications of simulated annealing selection metod. The maximum of ten replication The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The 'SA AVG' and 'SA MAX' test accuracy are increasing as the number of features used increases, while the 'Greedy' test accuracy may decrease. In addition, for k-nearest neighbor classifier, the feature selected by simulated annealing algorithm is better than greedy feature selection on average test accuracy as well as maximum test accuracy when the number of features is large.

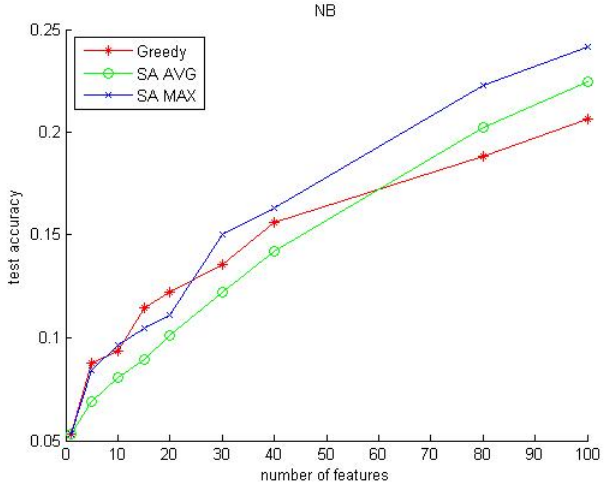

Fig. 5. Comparison of test accuracy between greedy and simulated annealing feature selection methods with document frequency (DF) statistic using Naive Bayes classifier. 'Greedy' denotes the test accuracy using the feature selected by greedy feature selection method. 'SA AVG' and 'SA MAX' denote the average and maximum test accuracy, respectively, of ten replications of simulated annealing selection metod. The maximum of ten replication The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The 'SA AVG' and 'SA MAX' test accuracy are increasing as the number of features used increases, while the 'Greedy' test accuracy may decrease. In addition, for Naive Bayes classifier, the feature selected by simulated annealing algorithm is better than greedy feature selection on average test accuracy as well as maximum test accuracy when the number of features is more than 80. But when number of features is small, feature selected by greedy is as good as the maximum accuracy of feature selected by simulated annealing algorithm.

Naive Bayes classifier. The test accuracy using feature selected by greedy algorithm, average test accuracy and maximum test accuracy of ten replications of simulated annealing algorithm are shown in figure 5. From the plot, we can see that the feature selected by simulated annealing algorithm is better than greedy feature selection on average test accuracy as well as maximum test accuracy when the number of features is more than 80. But when number of features is small, feature selected by greedy is as good as the maximum accuracy of feature selected by simulated annealing algorithm.

5) Comparison between Greedy and Simulated Annealing with SVM: In this section we compare the feature selected by greedy and simulated annealing algorithms using SVM classifier. The test accuracy using feature selected by greedy algorithm, average test accuracy and maximum test accuracy of ten replications of simulated annealing algorithm are shown in figure 6. From the plot, we can see that the feature selected by simulated annealing algorithm is better than greedy feature selection on maximum test accuracy when the number of features is more than 30, while on average test accuracy two algorithm have similar results. But when number of features is small, feature selected by greedy is better than the maximum accuracy of feature selected by simulated annealing algorithm.

6) Greedy Feature Selection with Mutual Information: We use greedy feature selection method to select 100 features which achieves highest conditional mutual information score. Then three classifiers (k-nearest neighbor, Naive Bayes and SVM) are used to test the feature selected. The test accuracy is shown in figure 7 where the classification is done using the first 1, 5, 10, 15, 20, 30 ,40, 80 and 100 features. The test accuracy is strictly increasing when number of features is increasing for three classifiers. In addition, SVM always outperforms Naive Bayes which beats k-nearest neighbor at most cases.

7) Greedy Feature Selection with χ^2 Statistic: We use greedy feature selection method to select 100 features which achieves highest conditional χ^2 statistic score. Then three classifiers (k-nearest neighbor, Naive Bayes and SVM) are used to test the feature selected. The test accuracy is shown in figure 8 where the classification is done using the first 1, 5, 10, 15, 20, 30 ,40, 80 and 100 features. The test accuracy is strictly increasing when number of features is increasing for three classifiers. In addition, SVM always outperforms Naive Bayes and k-nearest neighbor. When number of features is small, k-nearest neighbor is better than Naive Bayes, while when number of features is greater than 40, k-nearest neighbor losses to Naive Bayes.

D. Latent Features: PCA

In this section, we also explore an alternative method (Latent Features) to deal with dimensionality reduction problem in 20 newsgroups text classification. We first normalize the data and then use PCA to project the original feature space to a lower dimensional space using linear combination of original features. The scree plot and plot of the variance explained by the number of principal components are shown in figure 9.

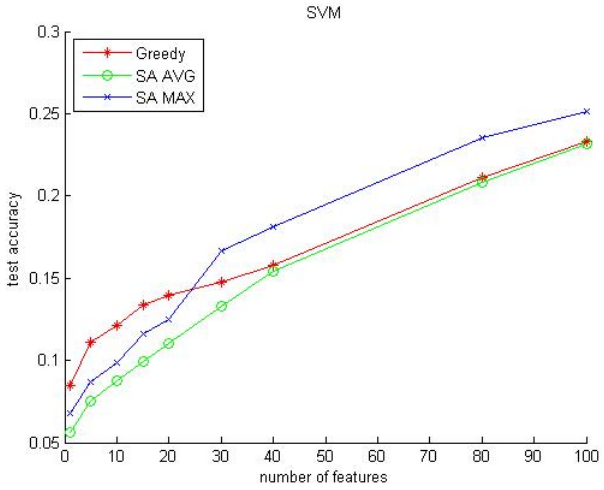

Fig. 6. Comparison of test accuracy between greedy and simulated annealing feature selection methods with document frequency (DF) statistic using SVM classifier. 'Greedy' denotes the test accuracy using the feature selected by greedy feature selection method. 'SA AVG' and 'SA MAX' denote the average and maximum test accuracy, respectively, of ten replications of simulated annealing selection metod. The maximum of ten replication The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The 'SA AVG' and 'SA MAX' test accuracy are increasing as the number of features used increases, while the 'Greedy' test accuracy may decrease. In addition, for SVM classifier, the feature selected by simulated annealing algorithm is better than greedy feature selection on maximum test accuracy when the number of features is more than 30, while on average test accuracy two algorithm have similar results. But when number of features is small, feature selected by greedy is better than the maximum accuracy of feature selected by simulated annealing algorithm.

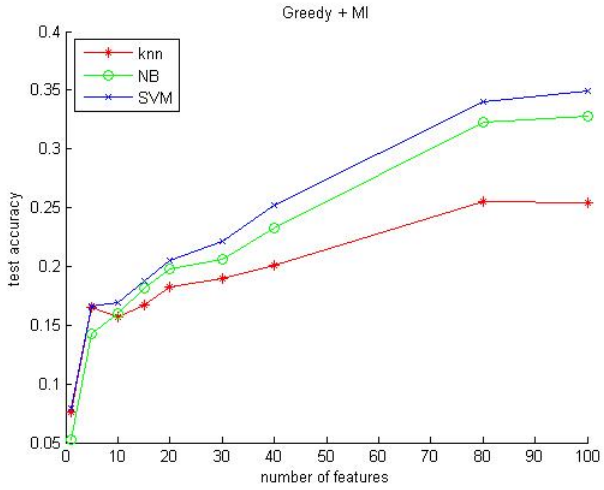

Fig. 7. Test accuracy of three classifiers, k-nearest neighbor, Naive Bayes and SVM. Feature used in classification is selected by greedy feature selection method and mutual information (MI) statistic. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The test accuracy is increasing as the number of features used increases in most cases. In addition, SVM always has the best accuracy and Naive Bayes follows, k-nearest neighbor is the worst.

we do the classification using the first 1, 5, 10, 15, 20, 30, 40, 80 and 100 principal components generated by PCA. The classifiers used are k-nearest neighbor and SVM. The reason that Naive Bayes classifier is not included is that after projection in PCA the value of certain principal components might be negative. The test accuracy is shown in figure

E. Comparison between Feature Selection and Latent Feature

In this section, we compare the results of features/principal components selected by different dimensionality reduction strategies. Since from previous results, we can see that SVM always has the best test accuracy compared to k-nearest neighbor and Naive Bayes classifiers, then in this section, we use the test accuracy of SVM to compare the performance of different dimensionality reduction strategies as shown in figure 11. The first dimensionality reduction method is the feature selection based on χ^2 statistic, which is provided by python standard library (scikit learn), denoted by 'STD + CHI'; the second is feature selection based on conditional score of χ^2 statistic as described in section II-C, denoted by 'Greedy + CHI'; the last on is PCA as shown in section III-D, denoted by 'PCA'.

IV. Conclusions

In this project, we implement the feature selection using greedy and simulated annealing algorithm based on conditional score of document frequency, mutual information and

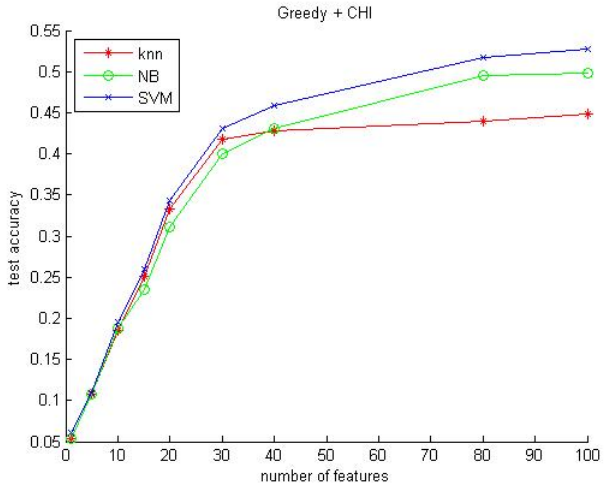

Fig. 8. Test accuracy of three classifiers, k-nearest neighbor, Naive Bayes and SVM. Feature used in classification is selected by greedy feature selection method and mutual information (MI) statistic. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 features. The test accuracy is increasing as the number of features used increases. In addition, SVM always has the best accuracy and Naive Bayes follows, k-nearest neighbor is the worst when number of features is greater than 40.

χ^2 statistic. Then we compare the test accuracy using three classifiers (k-nearest neighbor, Naive Bayes and SVM) and features selected by greedy and simulated annealing algorithm. In addition, we explore other dimensionality reduction technique: Latent Features (PCA).

From the results mentioned above, to answer the questions introduced at the beginning, we conclude that

1) With document frequency (DF), by using simulated annealing algorithm in feature selection instead of greedy algorithm, the test accuracy can be improved among all three classifiers. Hence, the method we proposed

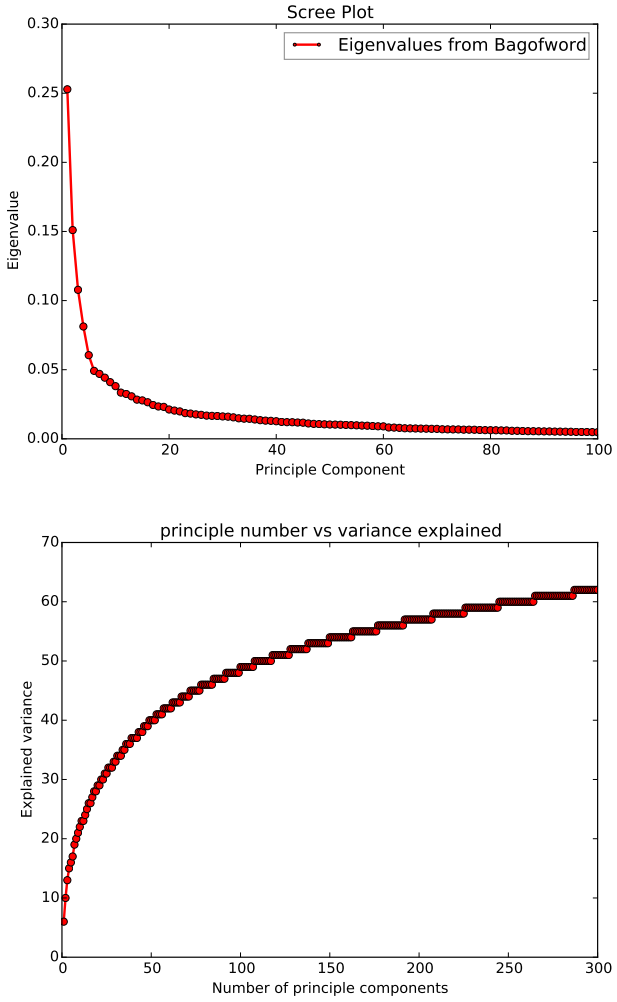

Fig. 9. Scree Plot and Variance Explained by the Number of Principal Components

is better than the existing method. In addition, our simulated annealing algorithm using mutual information and χ^2 statistic is still running due to the complexity of simulated annealing algorithm and the required several replications of simulated annealing algorithm, we are expecting to include the results in the final presentation if finished.

2) Although the test accuracy of 100 features for all classifiers are smaller than test accuracy using entire feature space which contains 129797 features, by using feature selection techniques, we can achieve 60% of test accuracy only by 0.1% of all features which is impressive. In addition, this can also prove that the validity and effectiveness of feature selection.

REFERENCES

[1] T. Joachims "Transductive inference for text classification using support vector machines," *International Conference on Machine Learning (ICML)* 200-209

[2] A. McCallum, K. Nigam "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization* 752, 41-48

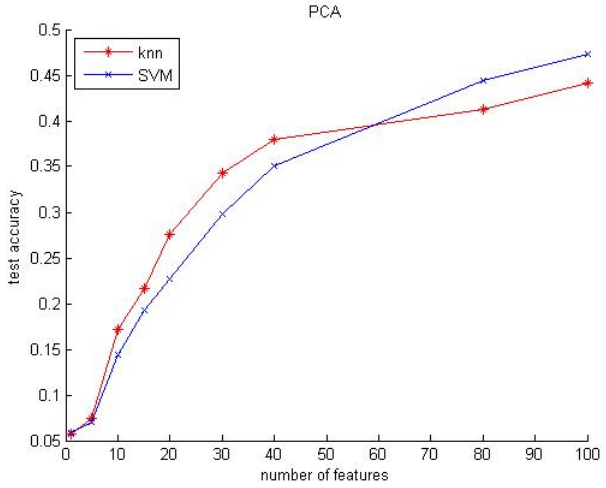

Fig. 10. Test accuracy of two classifiers, k-nearest neighbor and SVM. The plot shows the results using first 1, 5, 10, 15, 20, 30, 40, 80 and 100 principal components generated by PCA. The test accuracy is increasing as the number of principal components used increases. In addition, SVM has the better accuracy than k-nearest neighbor when number of principal components is greater than 80 while worse otherwise.

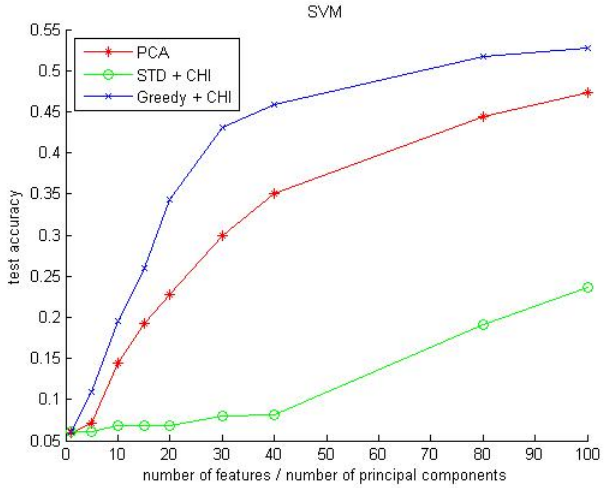

Fig. 11. Test accuracy of SVM using the features/principal components selected by three different dimensionality reduction strategy. PCA stands for latent features method which is a linear projection of original feature space. STD+CHI stands for feature selection method which is provided by python standard library based on χ^2 statistic. Greedy+CHI stands for feature selection method which selects features that achieves best conditional χ^2 statistic score.

[3] K. Nigam, A. McCallum, S. Thrun, T. Mitchell "Text classification from labeled and unlabeled documents using EM," *Machine learning* 39 (2-3), 103-134

[4] M. Rogati and Y. Yang "High-performing feature selection for text classification," *ACM CIKM* 2002.

[5] Y. Yang and J. O. Pedersen "A comparative study on feature selection in text categorization." *In International Conference on Machine Learning* 412420, 1997.

[6] J. Zhu, G. Bilbro and M.-Y. Chow "Phase Balancing using Simulated Annealing" *IEEE Transactions on Power Systems* vol. 14, no. 4, Nov. 1999, pp. 1508-1513.