Variance Based Classifier Comparison in Text Categorization

Atsuhiro Takasu and Kenro Aihara
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
E-Mail: {takasu,kenro.aihara}@nii.ac.jp

Text categorization is one of the key functions for utilizing vast amount of documents. It can be seen as a classification problem, which has been studied in pattern recognition and machine learning fields for a long time and several classification methods have been developed such as statistical classification, decision tree, support vector machines and so on. Many researchers applied those classification methods to text categorization and reported their performance (e.g., decision tree[3], Bayes classifier[2], support vector machine[1]). Yang conducted comprehensive study of comparison of text categorization and reported that k nearest neighbor and support vector machines works well for text categorization[4].

In the previous studies, classification methods were usually compared using single pair of training and test data However, classification method with more complex family of classifiers requires more training data and small training data may result in deriving unreliable classifier, that is, the performance of the derived classifier varies much depending on training data. Therefore, we need to take the size of training data into account when comparing and selecting a classification method. In this paper, we discuss how to select a classifier from those derived by various classification methods and how the size of training data affects the performance of the derived classifier.

In order to evaluate the reliability of classification method, we consider the variance of accuracy of derived classifier. We first construct a statistical model. In the text categorization, each document is usually represented with a feature vector that consists of weighted frequencies of terms. In the vector space model, document is a point in high dimensional feature space and a classifier separates the feature space into subspaces each of which is labeled with a category.

Let us consider the problem of classifying documents into c categories, and suppose we obtain a classifier which separate the feature space into m subspaces $s_1, s_2, \cdots, s_m$. In the case of Rocchio's classification method, the number m of the future subspaces is the number c of categories, while it is the number of leaves for decision trees. Let p_i denote occurrence probability, that is, the probability that a document vector is in subspace s_i. Notice that $\sum_{i=1}^{m} p_i = 1$ holds. Suppose the category of a subspace s_i is c_i, then the accuracy of s_i, denoted by α_i, is the probability that the category of document in s_i is c_i. The accuracy of the classifier is described as follows.

$$\sum_{i=1}^{m} p_i \alpha_i \tag{1}$$

Suppose that a classifier is learned from n training documents Let n_i denote the number of documents in subspace s_i, and a_i denote the number of correct documents in s_i. Then, using the empirical occurrence probability $\hat{p}_i = n_i/n$ and the empirical accuracy $\hat{\alpha}_i = a_i/n_i$, the empirical accuracy is denoted as

$$\sum_{i=1}^{m} \hat{p}_i \hat{\alpha}_i \tag{2}$$

Let us consider the variance of expression (2). Accuracies α_i and α_j are independent for any i, j $(i \neq j)$. Accuracy α_i and the occurrence probability p_j are independent for any i, j, too. The mean of $\hat{\alpha}_i$ is α. Therefore, the variance of the empirical accuracy is described as follows:

$$\begin{aligned} V[\sum_{i=1}^{m} \hat{p}_i \hat{\alpha}_i] &= \sum_{i=1}^{m} \sum_{j=1}^{m} E[\hat{\alpha}_i] E[\hat{\alpha}_j] C[\hat{p}_i, \hat{p}_j] \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j C[\hat{p}_i, \hat{p}_j] \end{aligned} \tag{3}$$

where E, V and C stand for expectation, variance and covariance, respectively.

By equation (3), we obtain the variance of empirical accuracy from the covariance of the empirical occurrence probabilities estimated from n training data. For this purpose, we consider Fisher's amount of information for the occurrence probabilities. In case of prior probabilities, the probability distribution function is multinomial distribution

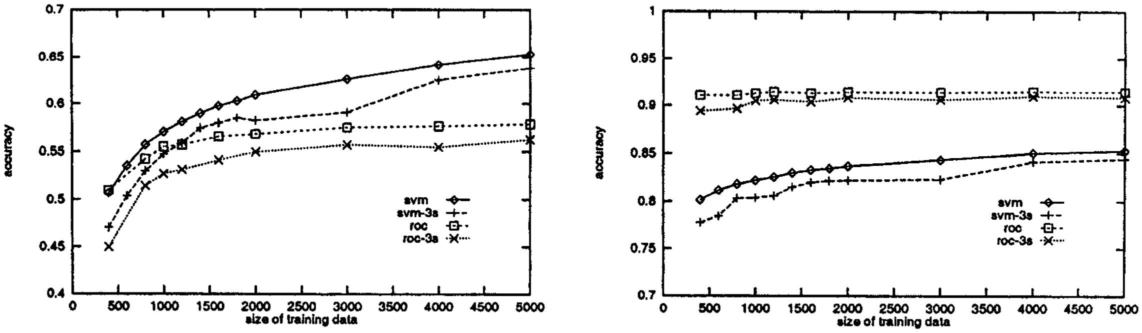
$$f(n_1, n_2, \cdots, n_m) = \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdot p_m^{n_m} .$$

Since $\sum_{i=1}^{m} p_i = 1$ holds, we replace p_m with $1 - \sum_{i=1}^{m-1} p_i$ and use the following Fisher's amount of information $I(\mathrm{P})$ for parameters $\mathrm{P} = (p_1, p_2, \cdots, p_{m-1})$:

$$\begin{aligned} I(\mathrm{P})_{ij} &= E[\frac{\partial \log f(\mathrm{P})}{\partial p_i} \frac{\partial \log f(\mathrm{P})}{\partial p_j}] \\ &= \begin{cases} 1/p_i + 1/p_m & i = j \\ 1/p_m & i \neq j \end{cases} \end{aligned} \tag{4}$$

This equation means that Since $\hat{P} = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{m-1})$ is maximum likelihood estimation(MLE) of P, the following convergence in distribution holds by the asymptotic efficiency of

SIGIR 2000 7/00 Athens, Greece



(a) 24 categorization

(b) 2 categorization

Figure 1: Accuracy and Variance of Text Categorization

MLE:

$$\sqrt{n}(\mathrm{P}-\hat{\mathrm{P}}) \xrightarrow{d} N(0, I(\mathrm{P})^{-1}), \quad (n \to \infty)$$

where $N(0, I(\mathrm{P})^{-1})$ stands for the normal distribution with the covariance matrix $I(\mathrm{P})^{-1}$. From equation (4), the inverse of Fisher's amount of information is

$$I(\mathrm{P})^{-1}_{ij} = \begin{cases} p_i(1-p_i) & i=j \\ -p_i p_j & i \neq j \end{cases} . \quad (5)$$

From equations (3) and (5) the following variance is derived.

$$V[\sum_{i=1}^{m} \hat{p}_i \hat{\alpha}_i] = \frac{1}{n}(\sum_{i=1}^{m} \alpha_i^2 p_i - \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j p_i p_j) \quad (6)$$

the variance is negligible when using sufficiently large training data, however we have to take the variance into account for small training data. We can derive lower bound of accuracy from this equation using, for example, Chebyshev inequality and select classifier based on the lower bound of the accuracy.

We made preliminary experiment on text categorization. In this experiment, we used a subset of NACSIS-IR database which consists of abstracts presented at academic conferences sponsored by 24 Japanese academic societies. We used these 24 societies as classes and defined the class of an abstract as the society sponsoring the conference which the abstract is presented at. This data set consists of 327,880 abstracts. Two kind of problems are considered, one is classifying a document into one of 24 classes, and the other is determining whether a document belongs to the class. We refer to the former problem as 24-categorization and the latter as binary categorization, respectively. In order to construct feature vectors of articles, 1,000 words are selected as features based on the information gain criterion[3]. Each document is represented with a term frequency vector of the selected 1,000 words weighted by tf-idf. We prepared 30 sets of training data for each size of training data ranging from 400 to 10,000. In order for each training data to contain at least one abstract from each class, we first chose one abstract randomly from each class, then chose remaining data from the database randomly. We prepared a test data containing 10,000 abstracts in the same way as the training data. Rocchio's method and SVM with Gausian radial basis function kernel were applied to those data sets. SVM^{light} was used to obtain SVM classifiers.

Figure 1(a) and (b) show the experimental result of 24-categorization and binary categorization problem for a society, respectively. In these figures, ◇ (svm) stands for the average accuracy of the classifiers derived by SVM for each size of training data. Let σ stand for the standard deviation of the accuracy. Then, + (svm-3s) plots the average accuracy subtracted by 3σ. Similarly, box (roc) and × (roc-3s) stand for the average accuracy of the classifiers derived by Rocchio's method and the one subtracted by 3σ, respectively. As these figures show, performance of classification varies depending on the problem. SVM outperforms Rocchio for 24-categorization problem, while Rocchio outperforms SVM for a binary categorization problem (figure 1 (b)). The variance affects the accuracy of classifiers for small training data especially in 24-categorization problem. In 24-categorization problem, classifiers derived by both the Rocchio's method and SVM decompose feature space into smaller regions than the binary categorization problem. From equation (6), larger number of regions tend to result in larger variance.

In this paper, we discussed the reliability of classifier from the variance point of view. In future, we will study quantitative relationship between the equation(6) and experimental results. We also plan to make experiment using other corpora.

References

[1] T. Joachims. "Text categorization with support vector machines: learning with many relevant features". In *Proc. of European Conference of Machine Learning*, 1998.

[2] D. Koller and M. Sahami. "Hierarchically classifying documents using very few words". In *Proc. of 14th International Conference on Machine Learning*, pp. 170–178, 1997.

[3] D. Lewis and M. Ringuette. "A comparison of two learning algorithms for text categorization". In *Proc. of 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81–92, 1994.

[4] Y. Yang and X. Liu. "A re-examination of text categorization methods". In *Proc. of 22th Annual International Conference ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–49, 1999.