China-USA Business Review, ISSN 1537-1514
September 2013, Vol. 12, No. 9, 911-917

Multi-valued Document Classification Based on Generalized Bradley-Terry Classifiers Utilizing Accuracy Information*

Tairiku Ogihara, Kenta Mikawa, Masayuki Goto
Waseda University, Tokyo, Japan
Gou Hosoya
Tokyo University of Science, Tokyo, Japan

Due to the development of computer network, a large amount of documents are treated in many fields. The number of digital document data stored in databases is enormous, accordingly it is difficult for analysts to read all documents and classify it by hand. Therefore, it is necessary to develop the technology of automatic document classification by using computers these days. From the above needs, many classifiers with good performance have been proposed, i.e., Relevance Vector Machine (RVM) and Support Vector Machine (SVM) that are known as good binary classifiers. For multi-valued document classification problems, it is known that a multi-valued classifier by combining several binary classifiers has a good performance. In this study, the method to construct an efficient combination of binary classifiers based on improving Generalized Bradley-Terry (GBT) model, which has high extensibility, is focused. This model is an expansion of Bradley-Terry (BT) model. Though the BT model has a limitation on combination of classes, the GBT model enables us to utilize any binary classifier which classifies into two arbitrary subsets in the class set. Generally, when several binary classifiers learn from the training dataset, there would be the difference of accuracy between these binary classifiers, due to the existence of categories that cannot be easily classified. However, the conventional method of multi-valued classification by GBT binary classifiers does not take the accuracy of each classifier into consideration. To avoid this problem, a new way of multi-valued classification method by considering each classifier's accuracy is proposed. The purpose of this study is to construct a good multi-valued classifier by calculating the accuracy of each classifier and utilizing it as the weight. In order to verify the effectiveness of the proposed method, the simulation experiment by using newspaper articles is conducted.

Keywords: Generalized Bradley-Terry (GBT) model, multi-valued classification, Relevance Vector Machine (RVM), document classification, the accuracy of each classifier, combination of binary classifiers

* KAKENHI, Grant-in-Aid for Scientific Research (C), No. 23510192.
Tairiku Ogihara, B.E. degree, Department of Industrial and Management Science, Faculty of Science and Engineering, Waseda University.
Kenta Mikawa, M.E. degree, Department of Industrial and Management Science, Faculty of Science and Engineering, Waseda University.
Masayuki Goto, Dr.E. degree, Department of Industrial and Management Science, Faculty of Science and Engineering, Waseda University.
Gou Hosoya, Dr.E. degree, Department of Management Science, Faculty of Engineering, Tokyo University of Science.
Correspondence concerning this article should be addressed to Tairiku Ogihara, 3-4-1 Okubo Shinjuku-ku Tokyo 169-8555, Japan. E-mail: i.w.g.p-land@akane.waseda.jp.

Introduction

Due to the development of computer networks, a large amount of digital documents have become to be able to be treated in every field. The number of document data stored in databases is enormous, accordingly, it is difficult for analysts to read all documents and classify it by hand. Therefore, the necessity of automatic document classification by using computers has risen these days. From the above needs, many methods by applying binary classifiers with good performance have been proposed. For example, Relevance Vector Machine (RVM) and Support Vector Machine (SVM) are known as good binary classifiers and these methods can be applied to the binary document classification problems (Tipping, 2001; Cortes & Vapnik, 1995). However, it is sometimes difficult to acquire a unique multi-valued classifier with both high performance and practical computational complexity. Thus, the formulation of the multi-valued classification problem by a combination of several binary classifiers has been proposed. The most basic idea is the one versus the rest method which prepares the binary classifiers to classify each category and others. Though the one versus the rest method is a very simple way, it is known that a combination of appropriate number of binary classifiers is more efficient than the one versus the rest method and a unique multi-valued classifier. Therefore, several works using binary classifiers for multi-valued classification problems have been studied (Rumelhart & McClelland, 1986; Quinlan, 1993). One of these methods by combining binary classifiers is the Error-Collecting Output Codes (ECOC) (Dietterich & Bakiri, 1995; Hastie & Tibshirani, 1998; Allwein, Schapire, & Singer, 2000). Another effective idea is to apply the Bradley-Terry (BT) model (Ikeda, 2010; Huang, Weng, & Lin, 2006). The BT model is a statistical model and has broad applications in many fields. The application of the BT model to the multi-valued classification is an attractive way and has a possibility to develop this research field.

In this study, the method to construct an efficient combination of binary classifiers based on Generalized Bradley-Terry (GBT) model (Huang et al., 2006) with high extensibility is focused. This model is an expansion of BT model (Bradley & Terry, 1952). Though the original BT model has a limitation on combination of classes, the GBT model enables us to utilize arbitrary binary classifier which divides into an arbitrary size of two subsets in the class set. Generally, when several binary classifiers are combined to implement multi-valued classification, there exists the difference of accuracy between binary classifiers, due to the existence of categories that cannot be easily classified. However, the conventional method of multi-valued classification by GBT binary classifiers does not take the accuracy of each classifier into consideration. Therefore, classifiers with bad performance can make the gross accuracy decrease. In short, a direct use of those bad classifiers leads to degrade the classification performance. To avoid this problem, a new way of multi-valued classification method by considering each classifier's accuracy is proposed. The purpose of this study is to develop the effective method by estimating the accuracy of each classifier and utilizing it as the weight of classifier. The proposed method gives the large weights to good classifiers with high accuracy and gives the small weights to bad performance classifiers. The improvement of performance by applying these weights in classification rule is expected. To verify the effectiveness of the proposed method, the simulation experiment by using newspaper articles is conducted.

Preliminaries

Multi-valued Classification Problems

Let the number of categories be K, and the set of categories be $C = \{c_1, \ldots, c_K\}$. x is an input vector which

has an unknown category. A classifier is adaptively modeled by a learning rule with the training data set. When a new test data is provided, its class is predicted by the trained classifier. The classification problem is defined as a prediction of the category $c \in C$ to which a new input *x* belongs. The classification problem is referred to as multi-valued classification problems in the case of $K \geq 3$, and is also referred to as binary classification problems in the case of $K = 2$.

Classifications by Using RVM

In this study, the RVM classifier is applied by performing soft decision to estimate the posterior probability of each category. The binary RVM classifier has a lot of same characteristics as the SVM which is a good binary classifier with high accuracy. The RVM was proposed by Tipping (2001), which is a sparse learning algorithm applied to regression and classification problems. Silva and Ribeiro (2006) talked about the application to the text classification problem. The RVM is similar to the SVM (Cortes & Vapnik, 1995) in many respects but is capable of expressing a fully probabilistic model.

First, a binary classification model ($K = 2$) by the RVM is explained. Let *x* be an input vector and $c \in \{c_1, c_2\}$ be a category label. A set of *N* training document samples is denoted by $\{x_n, t_n\}_{n=1}^{N}$, $t_n \in \{c_1, c_2\}$. The probability of category label *c* takes c_k ($k = 1, 2$) conditioned on *x* is expressed by using logistic regression as follows:

$$p(c = c_k \mid x) = \frac{1}{1 + \exp(-f_{RVM}(x))} \tag{1}$$

$$f_{RVM}(x) = \sum_{i=1, t_i = c_k}^{N} w_i K(x, x_i) \tag{2}$$

where $w_i \sim N(0, \alpha_i^{-1})$ and α_i^{-1} are obtained by maximizing a posterior probability of α, which is hyper parameter that controls the distribution of parameters. Moreover $K(...)$ denotes a kernel function which calculates inner product of two input data points mapped on a higher dimensional space, and w_i expresses a weight parameter. By maximizing a posterior probability, almost all α_i^{-1} becomes zero. x_i having non-zero w_i value the Relevance Vector (RV) is called. The decision function $f_{RVM}(x)$ is determined by these RVs. The RVM has several desirable properties and good performance in classification accuracy. On the other hand, the RVM needs to spend more times in learning phase compared to the SVM. If the RVM is performed for a model with *M* basis functions, the computational complexity of evaluating inverse matrix of size *M* takes $O(M^3)$ (Bishop, 2006).

For a multi-valued classification problem ($K \geq 3$), a probabilistic method of combining *G* linear models is used. The parameter α_i^{-1} is calculated in the same way as a case of two categories. It is a quite straightforward extension to the multi-valued classification problem, but there exists a disadvantage that the computational complexity for learning is K^3 times larger than that of the binary RVM (Bishop, 2006). Assume that the classifier *r* ($r = 1, \ldots, R$) classifies input vectors *x* into the two category sets, C_r^+ and C_r^-. Here, assume that $C_r^+, C_r^- \in C$, $C_r^+, C_r^- \neq 0$, $C_r^+ \cap C_r^- = 0$, and $C_r = C_r^+ \cap C_r^-$. The binary classifier is reduced to

the 1-vs-1 classifier when $|C_r^+|=|C_r^-|=1$. If $|C_r^+|=1$ and $|C_r^-|=K-1$, then the binary classifier is referred as to the 1-vs-the-rest-classifier. Let the performance of classifier r be $q_r(x)(0\le q_r(x)\le 1)$, then $q_r(x)$ and $1-q_r(x)$ can be seen as the estimations of $p(c\in C_r^-\,|\,c\in C_r, x)$ and $p(c\in C_r^+\,|\,c\in C_r, x)$, respectively.

Conventional Study

Basic Explanation of BT Model

The BT model has been widely applied in many areas, especially in sports statistics (Ikeda, 2010). The BT model is a model to quantify the strength of each player when many people play a lot of one-to-one match. Assume that there are K players and a player k (k = 1, …, K) has non-negative parameter p_k called strength. Probability that player k wins against player l (l = 1, …, K, $l \neq k$) by $p_k/(p_k + p_l)$ is denoted, and probability that player l wins against player k by $p_l/(p_k + p_l)$ is denoted. Let n_{kl} be the number of matches between players k and l. Let r_{kl} be a winning rate of player k against player l, and it is given by:

$$r_{kl} = \frac{\text{the number of mathches that player } k \text{ wins against player } l}{n_{kl}} \tag{3}$$

We assume there is no tie so the equation $r_{kl} + r_{lk} = 1$ holds. Let $F(p)$ be a log-likelihood function, and it is defined as follows:

$$F(p) = \sum_{k=1}^{K}\sum_{l=k+1}^{K}\left(r_{kl}\ln\frac{p_k}{p_k+p_l} + (1-r_{kl})\ln\frac{p_l}{p_k+p_l}\right) \tag{4}$$

where $p=(p_1, ..., p_K)$ denotes a strength vector. By maximizing $F(p)$ in equation (4) with respect to $\sum_{k=1}^{K} p_k = 1,\ p_k > 0,$ the maximum likelihood estimator $\hat{p}=(\hat{p}_1, ..., \hat{p}_K)$ is acquired. The players according to their strength from $\hat{p}$ are ranked.

Multi-valued Classifier by BT Model

Consider applying the BT model to multi-valued classification by combining binary classifiers. The log-likelihood function of BT model, $F_{BT}(p, x)$, is defined as follows:

$$F_{BT}(p,x) = \sum_{r=1}^{R}\left(q_r(x)\ln\frac{p_{k_r}}{p_{k_r}+p_{l_r}} + (1-q_r(x))\ln\frac{p_{l_r}}{p_{k_r}+p_{l_r}}\right) \tag{5}$$

where $q_r(x)$ denotes the output of the 1-vs-1 classifier, k_r and l_r denote the numbers of category in C_r^+ and C_r^-, respectively. Similar to the previous subsection, the maximum likelihood estimator $\hat{p}=(\hat{p}_1,...,\hat{p}_K), \sum_{k=1}^{K}\hat{p}_k = 1, \hat{p}_k > 0,$ is obtained by maximizing $F_{BT}(p, x)$ in equation (5) with respect to p. Then it can be treated $\hat{p}_k$ as $p(c_k\,|\,x)$. Since $\hat{p}_k$ represents an estimation of category c_k, the classification rule of x from $\hat{c} = \arg\max_{c_k\in C}\hat{p}_k$ is formulated.

Multi-valued Classification by GBT Model

It is difficult to attain high performance by multi-valued classifier using the BT model, since the number of learning data of 1-vs-1 classifier in this model is lower than that of 1-vs-the-rest-classifier. As a result, classification accuracy of 1-vs-1 classifier is not so good. However, the GBT model can use any combination of 1-vs-1 classifiers and 1-vs-the-rest-classifier (Huang et al., 2006). The Log-likelihood function of GBT model, $F_{GBT}(p, x)$, is defined as follows:

$$F_{GBT}(p,x) = \sum_{r=1}^{R} \left(q_r(x) \ln \frac{\sum_{c_k \in C_r^+} p_k}{\sum_{c_l \in C_r} p_l} + (1 - q_r(x)) \ln \frac{\sum_{c_k \in C_r^-} p_k}{\sum_{c_l \in C_r} p_l} \right) \tag{6}$$

Similar to the case of multi-valued classifier by BT model in the previous subsection, the maximum likelihood estimator $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_K)$, $\sum_{k=1}^{K} \hat{p}_k = 1, \hat{p}_k > 0,$ is obtained by maximizing $F_{GBT}(p, x)$ in equation (6) with respect to *p*. Then the classification rule of *x* from $\hat{c} = \arg\max_{c_k \in C} \hat{p}_k$ is formulated.

Proposed Method

In the conventional multi-valued classification based on the GBT model, this model assumes the same weight for all binary classifiers. Therefore, there is a possibility that a bad classifier affects the total decision by all classifiers. The proposed method takes the estimated accuracy of classifier in consideration. It is necessary to determine the weights of the classifiers. By applying these weights to the GBT model, the modified GBT (MGBT) model which is a new way of multi-valued classification method by considering each classifier's accuracy is proposed.

Method for Calculating the Weight of Each Classifier

Usually, the reliability of each classifier can be different because it classifies different categories and is learned individually. Therefore, in order to verify the reliability of each classifier, $q_r(x)$ [1] is regarded as confidence of classifier *r*. Classifier accuracy can be estimated from the output of training data. The accuracy of classifier *r* is derived as follows:

$$A_r = \sum_{\substack{x' \in \chi' \\ c(x') \in C_r^+}} q_r(x') \tag{7}$$

where χ' denotes a training data set, and $c(x')$ denotes the category of a training data x' ($x' \subset \chi'$). If the total confidence A_r is large, the classifier *r* is easy to classify. On the other hand if A_r is small, the classifier *r* is difficult to classify the training data. As a result, the weight of each classifier which corresponds to its accuracy is calculated. In order to normalize the confidence A_r based on its maximum value $\max_{r'=1}^{R} A_r$, the weight of the classifier *r*, ω_r is calculated as follows:

$$\omega_r = \frac{A_r}{\max_{r'=1}^{R} A_{r'}} \tag{8}$$

[1] In this research, $q_r(x)$ is treated as output of classifier, the performance of classifier, or confidence of classifier.

Classification Considering the Variation in Accuracies of Classifiers on GBT Model

By applying the weights accuracy of GBT classifiers, a new classification rule considering the variation in accuracies of classifiers is proposed on GBT model. Like equation (6), the log-likelihood function $F_{MGBT}(p,x)$ can be formulated as a follows:

$$F_{MGBT}(\boldsymbol{p},\boldsymbol{x}) = \sum_{r=1}^{R} \omega_r \left(q_r(\boldsymbol{x}) \ln \frac{\sum_{c_k \in C_r^+} p_k}{\sum_{c_l \in C_r} p_l} + (1 - q_r(\boldsymbol{x})) \ln \frac{\sum_{c_k \in C_r^-} p_k}{\sum_{c_l \in C_r} p_l} \right) \tag{9}$$

Similar to the case of multi-valued classifier by GBT model in the previous section, the maximum likelihood estimator $\hat{p} = (\hat{p}_1, ..., \hat{p}_K)$, $\sum_{k=1}^{K} \hat{p}_k = 1, \hat{p}_k > 0,$ is obtained by maximizing $F_{MGBT}(p,x)$ in equation (9) with respect to *p.* Then the classification rule of *x* is formulated by $\hat{c} = \arg\max_{c_k \in C} \hat{p}_k$.

Experiments

In order to verify the effectiveness of the proposed method, classification experiments by using newspaper articles and evaluating the classification accuracy are performed. In this study, the combination of 1-vs-the-rest-classifiers which are the most basic configuration of binary classifiers is focused.

Experimental Conditions

In this experiments, four categories (Economic, Social, Sports, and Entertainment) of the Mainichi Newspapers article published in 2000 are used. Every article belongs to only one category. For the numbers of training data 100, 300, and 500, the average values of classification accuracy are evaluated. Each experiment is repeated five times and the numbers of test data are equally 200. To estimate the parameters of the log-likelihood, the gradient method is applied. We use the word “frequency” as features of documents. The feature space is configured by words with more than 10 times appearance in whole documents. To evaluate the performance of the proposed method, we compare its accuracy with that of the original GBT model that does not take the accuracy of the classifier into account.

Result of Experiments

Figure 1 shows the results obtained by averaging the accuracies of experiments. In each experiment, the number of training data was set in three patterns: 100, 300, and 500. From the results, in the cases of 100 and 300 training data, the proposed method has statistically significantly higher classification accuracy than the conventional methods. In the case of 500 training data, there is no significant difference between two methods. It can be clarified that the proposed method is effective when using a small number of training data.

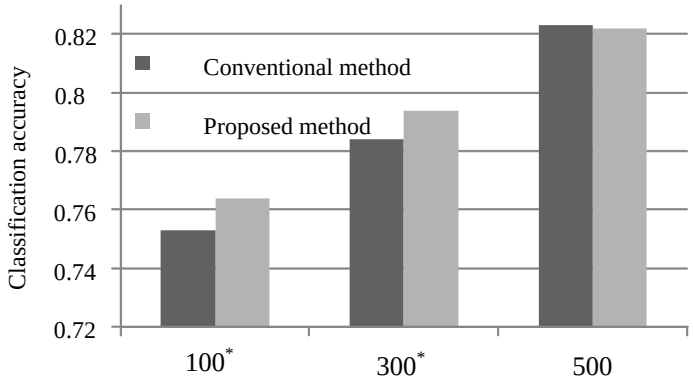

Figure 1. Classification accuracy by each number of training data (* shows that there is a 5% significant difference).

Discussion

The proposed method is an effective method when only a small number of training data are used in learning phase. If a large number of training data are used, the effectiveness of the proposed method is diminished because classification accuracy of each classifier can be improved and the effect of weighting is diminished. If we have to treat the more complex data structure, this characteristic is desirable in practice.

Conclusion and Future Work

In this study, the new multi-valued classification method by the GBT model considering the variation in the accuracy of each classifier is proposed and shows the effectiveness of the proposal by the experiments. The proposed method is effective in accuracy especially for the case only a small number of training data can be used for learning phase. The idea to use the information of accuracy of each classifier is useful for other methods with a combination of classifiers.

The proposed weighting method is based on a heuristic way. As future work, an examination of a method to investigate the optimum weight is remained.

References

Allwein, E. L., Schapire, R. E., & Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research, 1,* 113-141.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (pp. 56-67). New York: Springer-Verlag.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika, 39,* 324-345.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Journal of Machine Learning Research, 20,* 273-297.

Crammer, K., & Singer, Y. (2002). On the learnability and design of output codes for multiclass problems. *Machine Learning, 47,* 201-233.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research, 2,* 263-286.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics, 26,* 451-471.

Huang, T. K., Weng, R. C., & Lin, C. J. (2006). Generalized bradley-terry models and multi-class probability estimates. *Journal of Machine Learning Research, 7,* 85-115.

Ikeda, S. (2010). Combining binary machines for multiclass: Statistical model and parameter estimation. *The Institute of Statistical Mathematics, 58,* 157-166.

Lee, Y., Lin, Y., & Wahba, G. G. (2001). Multicategory support vector machines. Technical Report 1040, Department of Statistics, University of Madison, Wisconsin.

Quinlan, J. R. (1993). *Programs for machine learning*. San Mateo, C.A.: Morgan Kaufmann.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing explorations in the microstructure of cognition*. Cambridge, M.A.: MIT Press.

Silva, C., & Ribeiro, B. (2006). Scaling text classification with relevance vector machines (pp. 4186-4191). Proceedings from *SMC2006: IEEE International Conference on Systems, Man, and Cybernetics*.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1,* 211-244.