Web Page Classification Approach

P.Kameshwari, E.Salini Varma
Final Year Students
Department of Computer Science and Engineering,
Mahatma Gandhi Engineering College, India

Abstract

Textual document classification is one of the interesting areas of data mining. Textualdocuments may be arranged according to the topics or another characteristic (such as article type, writer, printing year etc.) some of the article consider only subject classification. Subject classification of documents based on two main ideas: the content-based method and the request-based method. Web page arrangement is onekind of textual document arrangement. Though, the text document presented in web pages is not similarin the meantime a web page can discuss correlated but dissimilar subjects. In consequence, results attained by a textual classifier are not as better as textual documents. Therefore, we need to improve the results of classifier using innovative technique.Firsttype of techniques that discourse this problem, by hidden the test setessential information to correct results,allocated by a textual classifier. In this article, discuss about a method that belongs to this category. Cross Training based Corrective approach (CTC) is new method for web page classification that acquiresdata from the test set in order to fixprimarily assigned by a text classifier on that test set. This technique can be tested using three basic classification algorithms: Support Vector Machine (SVM), Naïve Bayes (NB) and K- Nearest Neighbors (KNN), on four subdivisions of the Open Directory Project (ODP).

Keywords

Data mining, textual classifier, Cross Training based Corrective approach, Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN),

I. INTRODUCTION

Document classification *or* ***document categorization*** is a major difficulty in library science, information science and computer science. The task is to allocate a text to one or more modules or categories. This can be written by physically or algorithmically. The physical sorting of documents has generally been the authority of library science, whereas the algorithmic classification of documents is mainly in information science and computer science. The problems are overlying, however, and there is therefore interdisciplinary investigation on document classification.

The pamphlets can be classified by texts, images, music, etc. Every document classification possesses its superior classification problems. Document classification is the process of assigning classes to documents. The typicalmethod is to use a classifier such as Naïve Bayes, Support Vector Machine or K-Nearest Neighbors, to build a typical based on physically labelled documents. Then, existing test data are to the classifier, it uses that model to forecast a class for each item in the test set without considering other items in it.

In this paper, we suggest a post classification corrective approach called Cross Training Correction (CTC). This method is stimulated from the k-fold cross confirmation technique and uses the hidden information existing in the test set and the correlation between expected labels and attributes, in order to improve the result make category rectifications. In section 2, we discuss about CTC method in detail. In section 3, present experimental results.

II. PROPOSED METHOD

We propose a method to classify web page by using a corrective approach that can be applied in the context. This method improves results attained by a text classifier using the fundamental information hidden in the test set. It proceeds up to n iterations, where n is merely the number of test set's taken in our experiments. Let D= {(x, y)} be the dataset, where (x, y) is the pair of an instance, x is the features vector of the instance and y={0,1} is the class of the instance. First, dataset divided in to two distinct parts: a training set noted D_{Tr} and a test set noted D_T where D_{TR} D_T=D Then, we initiate the bootstrapping step, in which a classifier is trained using instances belonging to D_{Tr} to expect labels of instances belonging to D_T. In the adjustment step, which contains n iterations, we split D_T to n equal parts in order to obtain n distinct sets D_i, i=1, 2..., n where D_1 D_2 … D_n = D_T. After splitting, our method uses a cross-validation like sliding window of size w=t/n (t=$|D_T|$). In each iteration k (k=1, 2..., n), the sliding window contains web pages of Dk. Then the predict labels to train DT\D_k so that we can get some valuablebasic information from the

test set that was not in the training data. Then, the classifier dataset in $D_k$is applied (controlled by the sliding window), to accurate their classes. At last, the window travels to the next k-subset of D_T and the method is reiterated until the window scans all data in D_T. The entire correction step of the process is repeated up to convergence is obtained, or maximum number of iterations is reached. Figure 1 gives a summary of the corrective approach.

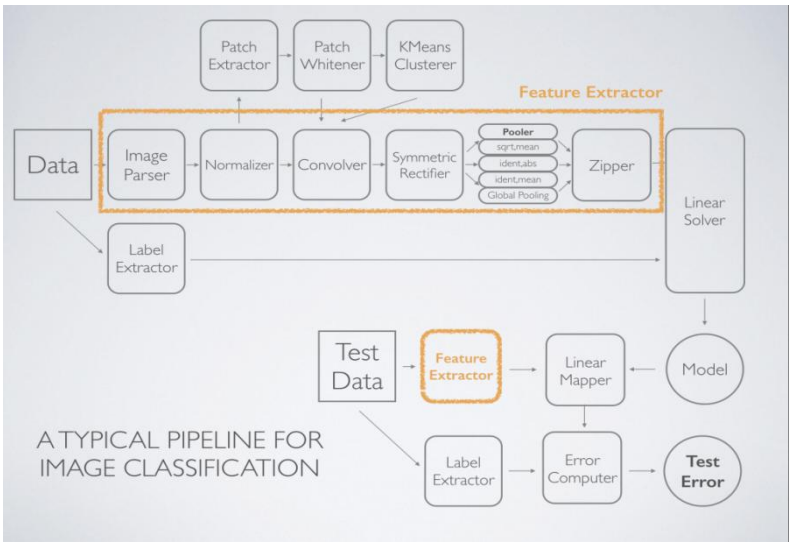


Fig.1 The correlative approach

III. DIFFERENT TYPES OF CLASSIFIERS

A) Support Vector Machine

The Support Vector Machines (SVM) is a great learning algorithm in text classification is performing well in this industry. It is based on the Structured Risk Maximization theory and main objective of this technique is to minimizing thesimplification error as a replacement of the experimental error on training data alone. Lin 2002 describes multiple versions of SVM and keerthi et al developed Sequential Minimal Optimization versionwhich is used in this session. Here we choose C=1 for the tolerance degree to errors. And used a linear kernel, which shows to be effective text categorization, where we have high feature vector dimension.

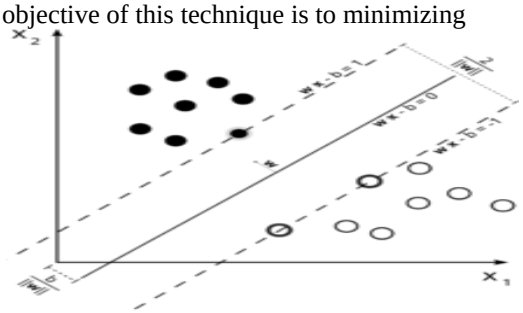
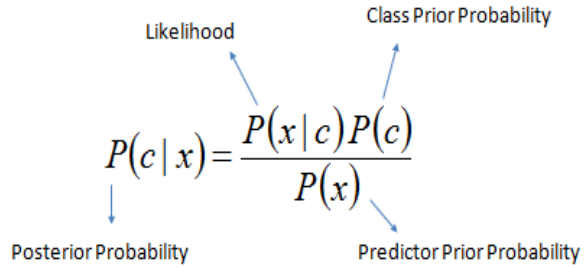


Fig. 2 Support Vector Machine

H_1 does not separate the classes.
H_2 does, but only with a small margin.
H_3 separates them with the maximum margin.

B) Naïve Bayes

In learning mechanisms, the **naive Bayes classifiers** are simple probabilistic classifiers put on Bayes' theorem with strong (naive) individuality expectations between the features. Naive Bayes has been studied widely since the 1950s.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naïve Bayes (NB) technique also used to test our approach. It is a simple and much known classification algorithm. The joint prospects of attributes and classes are used to estimate the probabilities of categories given a document, and make the assumption that features are conditionally independent of each other to make the computation of joint probabilities simple.

C) *K-Nearest Neighbors*

The K Nearest Neighbors (KNN) which is the utmost simple sorting algorithm. It is a kind of lazy learners or occurrence based learners that guess the category of an occurrence based on its K nearest examples in the feature space based on an inter-instance similarity. This algorithm used to stores the training samples directly and finds type of new instance but does not generate a model from training instances.The cross-validation for the training set needs to find applicable k parameter.

D) Datasets

This technique considers four binary problemsit can be simplyprolonged to multi-label classification problems by spread over "one against others" classification.Open Directory Project (ODP)used to select dataset which is anincrediblesource containing around 5 million webpages and is controlled into 765,282 groups and subgroups. The four binary classification tasks: "Male" vs. "Other" (1700 web pages), "Female" vs. "Other" (1690 web pages), "Kids" vs. "Other" (1782 web pages), and "Senior citizen" vs. "Other" 2013 web pages.This approach needs to train some web pages in content based classifier and some web pages to test the method. The flowing models consider one fold for training and remaining nine fold We used one fold for testing.

IV. RESULTS AND DISCUSSION

Based on three classifiers such as KNN, SVM and Naïve Bayes we consider almost all datasets from Table 1, which encompasses results obtained with a number of splits equal to 10. This demonstrates that the correlation between predicted labels and web pages attributes helps improve classification results. Unusually, our proposed method does not increase performances when using KNN on kid's dataset. Because of memory of KNN on senior citizen dataset is very low (0.365). This earns that huge instances belonging to senior citizen category, are categorized as other by KNN. Thus, classifier used for correction suffers from many noises created by wrong dependencies between labels and data.

Table 1. performance model

	Male			Female			Kids			Senior citiz		
	Accuracy	Memory	Fake	Accuracy	Memory	Fake	Accuracy	Memory	Fake	Accuracy	Memory	Fake
KNN	0.628	0.685	0.780	0.536	0.999	0.746	0.910	0.255	0.445	0.406	0.875	0.561
KNN+CTC	0.652	0.813	0.736	0.491	0.985	0.678	0.873	0.343	0.457	0.44	0.922	0.609
NB	0.902	0.710	0.805	0.87	0.835	0.403	0.910	0.79	0.701	0.782	0.878	0.819
NB+CTC	0.996	0.796	0.890	0.704	0.7	0.891	0.925	0.833	0.914	0.867	0.809	0.796

SVM	0.799	0.808	0.810	0.697	0.699	0.696	0.89	0.1007	0.943	0.889	0.981	0.837
SVM+CTC	0.897	0.873	0.879	0.761	0.874	0.701	0.987	0.85	0.963	0.909	0.9	0.998

V. CONCLUSION

In this paper,we consider Cross Training based Corrective approachthat helps improve text classifier results.This method used to identify the hidden information present in the test set to cooperatively adjust types of web pages. The above experiments illustrate that within asuitableexperimental setting, this approach progressesperformance of three classifiers: SVM, Naïve Bayes, and KNN. The textual classifier initially allocates the correlation between forecast labels and web pages that helps to regulating the classes. Results found after the original classification have an impact on the performance of our corrective approach.

REFERENCES

[1] Aha, D., et D. Kibler. 1991. « Instance-based learning algorithms ». Machine Learning 6: 37-66.

[2] Breiman, Leo. 1996. « Bagging Predictors ». Machine Learning 24 (2): 123-40.

[3] F. Mosteller, et J. W. Tukey. « Data Analysis, Including Statistics ». In Handbook ofSocial Psychology (G. Lindzey and E. Aronson, eds.), 2ᵉéd., 2:80-203. Addison-Wesley, Reading, MA.

[4] Freund, Yoav, et Robert E. Schapire. 1996. Experiments with a New Boosting Algorithm.

[5] Henderson, Lachlan. 2009. « Automated Text Classification in the DMOZ Hierarchy ».

[6] Jones, Karen Spärck. 1972. « A statistical interpretation of term specificity and its application in retrieval ». Journal of Documentation 28: 11-21.

[7] Liu, Yan, Zhenzhen Kou, Claudia Perlich, et Richard Lawrence. Intelligent System forWorkforce Classification.

[8] McCallum, Andrew, et Kamal Nigam. 1998. A comparison of event models for NaiveBayes text classification.

[9] Platt, John C. 1998. Sequential Minimal Optimization: A Fast Algorithm for TrainingSupport Vector Machines. ADVANCES IN KERNEL METHODS - SUPPORTVECTOR LEARNING.

[10] Rijsbergen, C. J. Van. 1979. Information Retrieval. 2nd éd.

[11] Wolpert, David H. 1992. « Stacked Generalization ». Neural Networks 5: 241-59.