A Feature Selection Method Based on Competition Winners Mechanism

Li Chengyan[1,a], Wang Xiaodong[1,a]
[1] Harbin University of Science & Technology, Harbin, Heilongjiang, 150080, China
[a] xiaodong19881002@126.com

Keywords. Text Categorization;Feature Selection;Information Gain

Abstract. Feature selection plays an important role in the field of text categorization. The traditional feature selection methods such as information gain(IG), the weight of evidence for text(WET) , χ^2 feature selection and so on are commonly applied in text categorization. However, the traditional feature selection methods are based on local features, there are many low information redundancies features was selected. In this paper a new feature selection algorithm called CWFS based on competition winners feature selection is proposed to solve the time consuming issue of classification as well as low accuracy. By removing uncorrelated or redundant features, the filter model only applies the top dozens of all features, thus the filter model can finish the classification with less time. Our concern is to reduce the dimension of the feature space. The classifiers with Naive Bayes and support vector machine have been used to run our experiment on TREC sets. The experimental results show that CWFS method can highly improve the quality of classification.

Introduction

In this increasingly digitally connected world, the dimensionality of data is also increasing. To handle this humongous amount of data, automated solutions so on. Data Mining is required. Data mining algorithms cause a huge computational cost which needs to be minimized. One of the primary steps for this is to ensure that only useful data reaches the data mining systems. The way to do this is via feature selection of data and subsequent filtering.

Data mining or knowledge discovery is a process of examining various data from different aspects and encapsulating it into more useful and profitable information by summarizing its inter-relationships. This could enhance decision making, thereby cutting costs and incrementing revenue. However, when the data in consideration is highly dimensionality, data mining algorithms incurs an exorbitant cost of computation. This is where feature selection algorithms come into picture, for they remove irrelevant features, cutting down the processing time drastically. There are mainly four types of feature selection methods namely filter, wrapper, embedded and hybrid method.

Text categorization is an important part of data mining. The task of text categorization is placing predefine indexing automatically, e-mail filtering, web browsing and personal information agents, text categorizations is an active and important research area where machine learning and information retrieve intersect. One major difficulty in text categorization is the large dimension of the feature space. Feature selection is a key step in text categorization, its results could pose a direct influence upon the classifier accuracy. The traditional feature selection methods are based on local features, there are many low information redundancies features was selected. CWFS is based on global features is proposed to improve this issue.

The rest of this paper is organized as follow: In section 2, related works were summarized. In section 3, we will discuss relevant feature selection methods. The new algorithm of feature selection will be discussed in Section 4. In the section 5, we will present experiment steps and results. Some conclusions and ideas for further research are described finally.

Related Works

Bellman[1], came with the so-called curse of dimensionality in 1957, though it seems logical that having a higher amount of information could lead to better results. Increase in dimensionality

increases the training time of algorithms too. Thereby removal of redundant and irrelevant features becomes vital. Blum and Langley[2], 1997 discuss the various factors that affects feature selection algorithm and divide them into four sectors namely: search strategy, search direction, evaluation criterion and stopping criterion. Guyon[3] and Elisseef[4] in 2003 refer to the recent developments in variable and feature selection which addressed the problem from the pragmatic point of view of improving the performance of predictors. They also recommended using a linear predictor and select variables in two alternate ways: (1) with a variable ranking method with a correlation coefficient or mutual information; (2) with a nested subset selection method performing forward or backward selection or with multiplicative update. Gavin Brown[5] in 2009 contributed a unifying theoretical understanding of filters and a natural way for deriving a space of possible ranking criteria based on mutual information between the class label and feature.

Common Feature Selection Methods

Feature selection is an important step in text categorization. Recently, a growing number of statistical classification methods and machine learning techniques have been applied in text categorization. The prevailing feature selection methods such as Information Gain(IG), Document Frequency(DF), χ^2 statistic(CHI), Expected cross entropy(ECE), weight of evidence for text(WET), Odds ratio(ODD). They are all explored in our experiment they will be introduced respectively as follow:

A. Informaiton gain

Information gain is commonly used as a feature goodness criterion in machine learning. It measures the amount of information obtained for category prediction by knowing the presence or absence of a feature in a document. We adopt m $\{c_i\}_{i=1}^m$ to denote the set of categories in the target spaces. The information gain of feature t is defined to be:

$$\mathrm{G}(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i)$$
$$+ p(t)\sum_{i=1}^{m} p(c_i|t)\log p(c_i|t)$$
$$+ p(\bar{t})\sum_{i=1}^{m} p(c_i|\bar{t})\log p(c_i|\bar{t})$$

In this function, c is the sample category, $p(t)$ is feature t of probability . $p(\bar{t})$ is feature t by the probability of non-appearing. $p(c_i|t)$ represents feature t appear in category c_i . $p(c_i|\bar{t})$ represents feature t non-appear in category c_i. For each feature the information gain is computed and those feature whose information gain is less than some predetermined threshold are removed from the feature space.

B. χ^2 statistic

$$\chi^2(t,c) = \frac{N\cdot(AD-CB)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)}$$

A, B, C, D represent quantity of document, showing in the following table, $N = A+B+C+D$

	c_i	$\overline{c_i}$
t	A	B
$\bar{t}$	C	D

The χ^2 statistic has a natural value of zero if t and c are independent. We computed for each category the χ^2 statistic between each feature in a training corpus and that category, and then combined the category-specific scores of each term into two scores:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} p_r(c_i)\chi^2(t,c_i)$$

$$\chi^2_{\max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\}$$

C. Expected cross entropy

$$ECE(t) = p(t)\sum_{i=1}^{m} p(c_i|t)\cdot\log_2\frac{p(c_i|t)}{p(c_i)}$$

The only difference in information gain is that expectations cross entropy method do not make features which do not exist.

D. Weight of evidence for text

$$WET(t) = p(t)\cdot\sum_{i=1}^{n} p(c_i)\cdot\left|\log_2\frac{odds(c_i|t)}{odds(c_i)}\right|$$

$$= p(t)\cdot\sum_{i=1}^{n} p(c_i)\cdot\left|\log_2\frac{p(c_i|t)(1-p(c_i))}{p(c_i)(1-p(c_i|t))}\right|$$

In this function,

$$odds(c_i|t) = \frac{p(c_i|t)}{1-p(c_i|t)} \quad \text{and} \quad odds(c_i) = \frac{p(c_i)}{1-p(c_i)}$$

is the evaluation function, and the later is a relative new evaluation function. It measures the difference between the probability of class and conditional probability of class for items, and take into account the t happen in the text.

E. Odds ratio

It is designed for two-class classifier,which is defined as follow:

$$OR(t) = \log_2\frac{odds(t|pos)}{odd(t|neg)}$$

$$= \log_2\frac{p(t|pos)(1-p(t|neg))}{p(t|neg)(1-p(t|pos))}$$

In this function, pos represent the positive instance and neg represents negative instances. ODD method does not treat all the class in the same as former evalution functions, it focus the value of target class, which make ODD method is especially suitable for two-class classifier. In two-class task, it is expected to distinguish the positive class but do not care about the negative class. But in real classification task, the negative instances usually occupy a percentage more than 90%, under this circumstance , the value of ODD has extra advantages than other information measures.

Proposed Approach

In this section, we will describe a new feature selection method based on a tournament system. Selection features by a global method such as global information gain may result in the selection of many redundant features with additional value.

The main idea of CWFS algorithm employ feature selection methods for each training sample set S feature scoring. So the training set to get a global score value for each feature. Each sample represents a single individual competition. For each sample, the features within are ranked by score value, top scoring feature in the sample is considered the winner of the competition. The winner of the competition features added to the classification model for classification.

$Algorithm: CWFS(S, G(), t)$
$in: Training\ set\ S\ of\ labeled\ examples\ with$
$boolean\ feature, global\ scoring\ method\ G(),$
$number\ of\ features\ t\ to\ select$
$out: weight\ vector(w)\ with\ new\ weights\ for$
$classification.$
$1: for\ each\ feature\ f\ present\ in\ S\ do:$
$2: compute\ G[f] = G(f)$
$3: winners[] = 0$
$4: for\ each\ X\ in\ S\ do:$
$5: select\ feature\ f\ with\ highest\ G[f]\ and\ X_f \neq 0$
$6: winners[f] = winners[f] + 1$
$7: t\ top - scoring\ features\ f\ in\ winners[]$

Figure 1. Competition winners feature selection algorithm pseudo-code

Steps 1-3, Sample set of each feature is calculated global score. Steps 4-6, each feature is ranked by global score. Finally, according to Step 7 selected t top-scoring features.

II. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental data sets

Use Trec05P-1, Trec07P, Ceas08-1 as experimental data sets relevant information is as follow.

corpus	language	ham	spam	total
Trec05p-1	English	39,399	52,790	92,189
Trec07p	English	25,220	50,199	75,419
Ceas08-1	English	27,126	110,579	137,705

Figure 2. Data set details

TREC is international abbreviation of the English text information retrieval conference. Professor Gordon V.Cormack Waterloo University organization, TREC spam filtering contest's goal is to provide a platform for global spam researchers and organizations. TREC2007 basic public data sets for data collection in English.

CEAS Spam-Filter Live Challenge competition is filtered by one of the world's leading anti-spam online information sessions organization, all filters in real-time, regular mail and spam stream test. Therefore, this data set compared to the previous one, the more focused and mail to simulate the real environment reception. This data set provides its greatest feature is the number of dynamic training feedback to simulate the real environment spam filtering, which show why the game is to become a real-time Challenge (Live-challenge).

B. performance evaluation

In this paper, the rate of false positive samples of patients with false positive rate and negative samples are sample as the evaluation criteria, the last generation 1-ROCA value.

Predicted / true	+	-
+	True positive(TP)	False negative(FN)
-	False positive(FP)	True negative(TN)

Figure 3. Evaluation Description

In the table above, "+"represents positive samples. "-" represents negative ones. TP positive samples were correctly judged as positive samples. FN positive samples were wrongly judged as negative samples. FP negative samples were wrongly judged as positive samples.TN negative samples were correctly judged as negative samples. The actual positive samples P=TP+FN. The actual negative samples N=FP+TN.

Positive samples of misjudgment rate,PMR%

$$PMR\% = \frac{FN}{P} \times 100\%$$

Negative samples of misjudgment rate,NMR%

$$NMR\% = \frac{FP}{N} \times 100\%$$

PMR% and NMR% are smaller values that indicate lower false positive rate of classification model. The better the accuracy of the model. Therefore, The smaller (1-ROCA)% the value, the better the model show that the performance classification.

C. Experimental results

There are two purposes in the experiments. One is to compare the performance of our algorithm with traditional algorism such as IG , χ^2 statistic and so on. The other is to compare the performance of classifier in different number of features.

algorithm	Number of features			
	64	128	256	512
IG	0.4745	0.6032	0.1179	0.0132
ODD	0.0186	0.0039	0.0027	0.0010
χ^2 statistic	0.5087	0.2500	0.0358	0.0606
ECE	0.3910	0.0595	0.0394	0.0071
WET	0.1129	0.0619	0.0124	0.0017

Figure 4. Comparison of 1-ROCA%

According to the figure 4 above, we can find that same number of feature and the classifier, result of ODD and WET feature selection is better than any other feature selection algorithm. ODD and WET selected feature has a high amount of information. The real matter is that IG have to that the feature does not appear.

algorithm	*Number of features*			
	64	*128*	*256*	*512*
IG	*2.8750*	*3.7528*	*3.9069*	*6.9324*
ODD	*4.9398*	*5.2602*	*6.1979*	*7.8560*
χ^2 *statistic*	*2.5887*	*8.6226*	*9.5185*	*10.4921*
ECE	*11.3452*	*13.4176*	*18.1557*	*23.0969*
WET	*11.3452*	*14.9731*	*16.4289*	*21.9125*

Figure 5. Comparison of Classification time

From the figure 5 above, we can find that same number of feature and the classifier, classification time of IG and χ^2 statistic feature selection is obviously better than that of others feature selection algorithm.

Conclusions

This paper focus on feature selection in text categorization applications. The traditional feature selection method based on local features. Therefore many redundancies was selected inevitably. To solve this problem, CWFS is proposed. CWFS based on global features. Finally, experimental results show CWFS improve the quality of classification. Our future work is mainly to improve the effectiveness of feature selection. This approach has used in a short text such as Twitter and web pages.

References

[1] I. Guyon, A. Elisseeff. “An Intorduction to Variable and Feature Selection” Journal of Machine Learning Research, 2003, 13:1157-1182.

[2] I. Guyon, “Gene Selection for Cancer Classification using Support Vector Machines”. Machine Learning, 2002, 46:389-422.

[3] Yan Xu, “A Comparative Study on Feature Selection in Chinese Spam Filtering,” IEEE, 2012.

[4] Yu L, Liu H. Nichols. “Efficient Feature Selection via Analysis of Relevance and Redundancy”. Journal of Machine Learning Research, 2004, 5:1205-1224.

[5] Xiang Zhang, Mingquan Zhou. “A Combined Feature Selection Method for Chinese Text Categorization.” IEEE, 2012.

[6] Samina Khalid, Tehmina Khalil, “A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning.” Science and Information Conference,2014, 8:27-29.