Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization

基于词频和文本分类分布的卡方统计特征选择

ABSTRACT

Text categorization (**TC**) becomes the key technology to find relevant and timely information from a volume of digital documents, and feature selection techniques are proposed to overcome the high dimensionality（高维度）which causes the high computational complexity and low accuracy in TC tasks（这导致了文本分类任务中的 计算复杂度和低准确性）. Chi-square statistics (CHI) is one of the most efficient feature selection methods; however, it has two weaknesses. (1) It is document frequency based, and only counts whether the term occurs or not. Actually, high-frequency term occurring in few documents is often regarded as a discriminator in corpus(语料库中的鉴别器). (2) It does not consider the term distribution. A term has more discriminating power(更多的区分能力) for a specific category when its difference in degree of distribution is lower（分布的差异程度较小时）. In this paper, we propose a modified CHI feature selection approach which is called **term frequency and distribution based CHI** to overcome these weaknesses. We use sample variance(样本方差) to calculate the term distribution, and improve the classic（经典的）CHI with maximum term frequency. Extensive（广泛的） and comparative experiments on three corpora show that the proposed approach is comparable to（比得上）the classic feature selection methods in terms of macro-F1 and micro-F1.

1. INTRODUCTION

Text categorization (TC) is defined as assigning new unlabelled documents to a set of pre-defined categories based on classification patterns [1,2,3].(文本分类是将新的未标签文档分配给一组基于分类模式的预定义分类)。 The volumes of digital documents available online are growing exponentially（以指数方式）. TC becomes a key technology to find relevant and timely information from these documents for many applications [4], such as customer relationship management [5], spam email filtering [6,7], web page classification [8], text sentinel classification [9]（文本标记分类）, software bug classification [10], etc.

TC is naturally treated as a supervised（受监督的）learning problem, and several algorithms from machine learning (ML) approaches have been used as TC classifiers in the past years, such as k-nearest neighbour (kNN) [11], support vector machines [12], and Naive Bayes [13,14]. In recent years, many other classifiers are proposed by researchers. Zhang et al. （张等人）[15] propose a novel projected prototype based classifier（全新的基于分类的预测模型）, in which a document category is represented by a set of prototypes, each assembling a representative for the documents in a subclass and its corresponding term subspace. Nguyen et al. [16] propose an improved centroid-based classifier（质心分类算法）which uses the precise term-class distribution properties（精确地词类分布属性）instead of the presence or absence of terms in classes（类别中词的是否存在）.

A major problem which makes TC different from other classification tasks is the high dimensionality of the feature space due to(由于) a large number of terms. This problem increases the computational complexity of ML methods used for TC and brings about inefficiency as well as low accuracy due to redundant（冗余） or irrelevant terms（不相关的词） in the feature space [3,11,17,18]. Therefore, feature selection (FS) techniques are proposed to reduce the high dimensionality under the premise（前提） of guaranteeing the performance of classifiers.

From document frequency perspective（角度、视角）, the classical methods almost use document frequency (DF) sufficiently, but we also cannot ignore the impact of term frequency (TF) [19]. For instance, two words having TFs of 10 and 100, respectively（分别的）, and DFs are both 10, which means that we are unable to judge their relative importance according to DF; on the other hand, TF considers such information which may be useful in thein selection of important features.

In this paper, we propose a modified chi-square statistics（卡方统计） (CHI) FS approach called term frequency and distribution based CHI (TFDCHI)(称为基于词频和分布的CHI). Then, we do experimental verification（进行实验验证） with kNN classifiers on common corpora(语料库). The rest of this paper is organized as follows: in Section 2, we describe the related work about FS, and analyse the drawbacks of four classic FS methods; in Section 3, we propose our modified chi-square statistics FS method based on term frequency and the term distribution; in Section 4, we describe the data-set, classifier（分类器）, and performance measures（性能度量） used in our experiments; Section 5 presents the experimental results and shows the effectiveness（有效性）of our new approach; and finally, we conclude this paper with future work. （最后，我们以未来的工作来总结这篇论文。）

2. RELATED WORK

FS is a process that selects a subset from the original feature set according to some criteria(标准) of feature importance [3,20]. There are two major ways of viewing FS [21]. The first one is wrapper（包装器）approach that selects the term subset using the evaluation function which act as a wrapper around the classifier algorithm, and these features will be used on the same classifier algorithm [22]; the other is the filter approach that selects the feature subset from the original feature space using one evaluation function which is independent to the classifier algorithm [22]. （第一个是包装器（wrapper）方法，它使用评估函数选择术语子集，该函数充当分类器算法的包装器，并且这些特征将用于相同的分类器算法[22]; 另一种是使用一个评估函数从原始特征空间中选择特征子集的过滤方法，该评估函数独立于分类器算法[22]）.Since the filter FS approach is simple and efficient, it has been widely used in the TC, such DF [11,23,24], CHI [11,23,24], information gain (IG) [11,23,24], mutual information (MI) (交互信息)[11,25], and expected cross-entropy (ECE)（预期的交叉熵） [26]. The IG and CHI are the two most efficient feature selections, and DF is comparable with the performance of IG and CHI [11]. In this paper, our new approach is also based on **filter selection**. We will give detailed definitions on these classical methods. Formally（形式上）, for a specific category, we give the following definition used in [24,27]:

a is the number of documents with term and belongs to category

b is the number of documents with term and do not belongs to category

c is the number of documents without term and belongs to category

d is the number of documents without term and do not belongs to category

Thus, N is the number of documents in training set, N = a + b + c + d.

|  |  |  |
| --- | --- | --- |
|  | 属于C类 | 不属于C类 |
| 包含词t | a | b |
| 不包含词t | c | d |

2.1 The Basic Principle of Chi-square Statistic and Improvement

The CHI is used to measure the lack of independence between ti and Ck, and the basic principle of chi-square is the thing which determines whether the hypothesis is true by observing the deviation of the actual value and the theoretical value. The specific steps are given as follows:

1. Assume that two variables are independent.

2. Observe the deviation of the actual value and the theoretical value.

3. According to the deviation, decide to accept the original assumption or the alternative hypothesis.

In step 3, if the deviation is small enough, we can argue that this error is caused by imprecise measuring methods or accident. Actually, two variables are independent; therefore, we accept the original assumption. However, if the deviation is big enough, we consider that it is unlikely to be accidental or caused by inaccurate measurement, and then we accept the alternative hypothesis.

The CHI formula is used to calculate the deviation of the actual value and the theoretical value. Assume that a term ti and a category Ck are independent, and the related variables are shown in Table 1.

Because of the independence of ti and Ck, the term ti should occur equiprobably in all documents, so this probability is defined as follows:

【公式一】

According to Table 1, whether a document containing a term ti belongs to category Ck has four circumstances. The number of documents that belongs to category Ck are a+c; therefore, in these documents, the number of documents where term ti should occur is given by

【公式二】

Then, according to Eq. (2), we can calculate the deviation of the actual value and the theoretical value for the documents with term ti which belongs to category Ck, the formula is shown as follows:

【公式三】

Similarly, the deviation of the remaining circumstances can be calculated and indicated as D2;D3; andD4. Finally, the deviation of term ti and category Ck is calculate by

【公式四】

This is the CHI formula, and then we simplify Eq. (4) and obtain the formula as follows:

【公式五】

If 卡方=0, the term ti and the category Ck are independent; therefore, the term ti does not contain any category information. Otherwise, the greater the value of the卡方, the more category information the feature ti owns.

The score of term ti in text collection is obtained by averaging or maximizing the category-specific scores:

【公式六】

【公式七】