Expert Systems with Applications 39 (2012) 4760–4768

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Comparison of term frequency and document frequency based feature selection metrics in text categorization

Nouman Azam *, JingTao Yao

Department of Computer Science, University of Regina, Regina, SK, Canada S4S 0A2

ARTICLE INFO

Keywords:
Text categorization
Feature selection metrics
Term frequency
Document frequency

ABSTRACT

Text categorization plays an important role in applications where information is filtered, monitored, personalized, categorized, organized or searched. Feature selection remains as an effective and efficient technique in text categorization. Feature selection metrics are commonly based on term frequency or document frequency of a word. We focus on relative importance of these frequencies for feature selection metrics. The document frequency based metrics of discriminative power measure and GINI index were examined with term frequency for this purpose. The metrics were compared and analyzed on Reuters 21,578 dataset. Experimental results revealed that the term frequency based metrics may be useful especially for smaller feature sets. Two characteristics of term frequency based metrics were observed by analyzing the scatter of features among classes and the rate at which information in data was covered. These characteristics may contribute toward their superior performance for smaller feature sets.

1. Introduction

The volume of digital documents available online is growing exponentially as a result of increased usage of the internet. Finding relevant and in time information from these documents are important for many applications. Automated text categorization is the key technology for this task (Shang et al., 2007). It has been utilized in many application areas such as, customer relationship management (Coussement & Van den Poel, 2008), spam email filtering (Sakkis et al., 2003; Zhou, Yao, & Luo, 2010), web page classification (Qi & Davison, 2009), text sentinel classification (Wang, Li, Song, Wei, & Li, 2011) and astronomy (Kou, Napoli, & Toussaint, 2005).

A moderate size text collection usually contains tens of thousands of features (Genkin, David, & Madigan, 2007; Yang & Pedersen, 1997). The commonly used 'bag-of-words' representation for text documents where each word is treated as a feature, results in high dimensionality (Sebastiani, 2002). Feature selection is among the possible solutions in such situations for making the learning task efficient.

Feature selection is an active research area in many fields such as, data mining, machine learning and rough sets (Liang, Wang, & Yao, 2007; Piramuthu, 2004; Yao & Zhao, 2008; Yao, Zhao, & Wang, 2008). Feature selection may be defined as the process of selecting most important features (Azam & Yao, 2011; Yao & Zhang, 2005). The process typically involves certain metrics that is used for finding utilities or importance level of features. Feature selection can help in efficient utilization of resources for large scale problems (Forman, 2003).

Existing feature selection methods in text categorization are either based on term frequency (López, Jiménez-Salazar, & Pinto, 2007; Moyotl-Hernández & Jiménez-Salazar, 2005; Tang, Shepherd, Milios, & Heywood, 2005) or document frequency (Forman, 2003; Lee & Lee, 2006; Mladenic & Grobelnik, 1999; Ogura, Amano, & Kondo, 2009; Yang & Pedersen, 1997). Term frequency is the number of times a particular word appeared in a document while document frequency is the count of documents containing that word. Term frequency may be considered as relatively more important, since document frequency is based on binary value of a term presence or absence in a document and it ignores the actual contribution of a word within a document. For instance, two words having term frequencies of 10 and 100, respectively, in a document will have same document frequency of 1. This means that we are unable to judge their relative importance for a document. Term frequency on the other hand considers such information which may be useful in selection of important features.

The above argument leads us to an interesting issue. If term frequency is relatively more important compared to document frequency, the same may be true for feature selection metrics defined with them. In other words, feature selection metrics defined with term frequency may perform better than those defined with document frequency. We will examine this issue in this article. In particular, we will compare term frequency and document frequency in a given feature selection metric for finding the more useful one. Two approaches may be utilized for this task.

* Corresponding author.
E-mail addresses: azam200n@cs.uregina.ca (N. Azam), jtyao@cs.uregina.ca (J.T. Yao).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.160

- Revising term frequency based feature selection metrics with document frequency, or,
- revising document frequency based feature selection metrics with term frequency.

We adopt the second approach in this research. Furthermore, the research is limited to recently proposed document frequency based metrics of discriminative power measure (Chen, Lee, & Chang, 2009) and GINI index (Shang et al., 2007). The metrics were revised with term frequency for investigating the effectiveness of the two types of frequencies.

2. Feature selection

Feature selection is a process which selects a subset of features that is considered as important. Such selection can help in building faster, cost effective and accurate models for data processing (Saeys, Inza, & Larranaga, 2007). A typical feature selection process consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and results validation (Liu & Yu, 2005). Subset generation is a searching procedure that generates a candidate feature subset for evaluation. The search for a subset may start at a full, empty or random feature set. The generated subset is then evaluated with an evaluation criterion which determines the goodness of the subset. The subset is compared with the best subset previously generated. The process of subset generation and evaluation are repeated until given stopping criteria is reached. Finally, the selected feature subset is validated with different tests using artificial or real world data.

Feature selection algorithms designed with different evaluation criteria fall broadly into three categories, namely, filter, wrapper and hybrid models. A filter model uses general characteristics of data for evaluating and selecting features and is independent of a mining algorithm. Evaluating features with a wrapper model requires a predetermined mining algorithm with its performance used as an evaluation criterion. A hybrid model combines the two models in a unified framework. Feature selection methods based on a wrapper model are mostly not suitable in large scale problems like text categorization (Forman, 2003). Majority of feature selection methods are based on a filter model which evaluates each feature independently (Forman, 2003). Feature selection algorithms use certain metrics to assign scores to features in such cases. Feature selection methods in text categorization are also sometimes refereed to as feature selection metrics (Forman, 2003).

Feature selection metrics can be defined mathematically by considering data points of the form $((x^1,y^1),(x^2,y^2),(x^3,y^3),\ldots,(x^m,y^m))$, where y^is are the class labels associated with each instance x^i. Each x^i is represented by a vector in D dimensions $(x_1^i,x_2^i,x_3^i,\ldots,x_D^i)$. The features used to define data points may be presented as a set $F=(f_1,f_2,f_3,\ldots,f_D)$. Feature selection metrics generate a score against each feature in F. A set $F' \subseteq F$ with $|F'| \leqslant |F|$ is generated using these scores. The set F' (i.e. the selected features) is commonly based on a threshold value or some predefined number of top scoring features.

3. Term frequency in feature selection metrics

We now elaborate the importance of term frequency for feature selection metrics. The metrics of discriminative power measure (*DPM*) and GINI index (*GINI*) are considered for this purpose. The *DPM* metric was proposed by Chen et al. (2009). It was very useful in reducing the feature set, for example, from thousands to hundreds of features with less than 5 percent of decreasing test accuracy. *DPM* was reported to have interesting properties on emphasizing classification in parallel and selection of both positive and negative features. Interested reader may refer to Chen et al. (2009) for more details on these properties. The *GINI* feature selection metric was proposed by Shang et al. (2007). It was based on the theory of GINI index which was previously used in decision trees for splitting attributes (Breiman, Friedman, Stone, & Olshen, 1984). The comparisons of *GINI* with several other metrics suggest that it was a useful metric, involving simpler computations (Shang et al., 2007).

The two metrics may be defined mathematically using the following notations. Let w be any word, we define its presence and absence in category i as follows.

A_i	number of documents with word w and belong to category i
B_i	number of documents with word w and do not belong to category i
C_i	number of documents without word w and belong to category i
D_i	number of documents without word w and do not belong to category i

The above notations may be used to define the total number of documents as $N = A_i + B_i + C_i + D_i$ and the total number of documents in category i as $M_i = A_i + C_i$. The *DPM* and *GINI* metrics for a word, w, are defined as follows.

$$DPM(w) = \sum_{i=1}^{m} \left| \frac{A_i}{M_i} - \frac{B_i}{N - M_i} \right| \quad (1)$$

$$GINI(w) = \sum_{i=1}^{m} \left(\frac{A_i}{M_i} \right)^2 \left(\frac{A_i}{A_i + B_i} \right)^2 \quad (2)$$

The documents belonging to a particular category i are referred to as positive documents while those not belonging to category i as negative documents. The fraction A_i/M_i in above equations may be understood as a word's probability given positive documents, i.e. its occurrences in documents of ith category, divided by total documents in ith category. In the same way, $B_i/(N - M_i)$ may be understood as a word's probability given negative documents. We may interpret the *DPM* and *GINI* according to these definitions. The *DPM* for a word is the absolute difference of a word's probability given positive and word's probability given negative documents. The category values are summed up to get the final *DPM* score. The *GINI* may be considered as the square of word's probability given positive documents weighted by the square of the word's probability given its entire occurrences (i.e. $A_i/(A_i + B_i)$). The category values are summed up to get *GINI* score for a word.

The *DPM* and *GINI* are dependent on A_i and B_i values as N and M_i are independent of a word frequencies. Since the values of A_i and B_i are document frequencies, we suspect their suitability in this case. We will consider some demonstrative examples for illustrating shortcomings that may result from using document frequency in these metrics. We make a couple of cases for this purpose. A sample database of documents of Table 1 will be considered.

Case 1. Word occurrences limited to single category: the document and term frequencies for three words are shown under respective categories in Table 2. We first consider the words *Hello* and *Price*. These words have same doc-

Table 1
Dataset of documents for demonstrative examples.

	Cat. A	Cat. B	Cat. C
Number of documents	40	30	20