Text Feature Weight Calculation Based on Class Information and Feature Entropy (Postprint)

Authors: Alimjan Aisha, Yin Xiaoyu, Kurban Wubuli, Li Zhe

Date: 2018-08-13T00:00:00+00:00

Abstract

Text vectorization is fundamental to text classification, and feature weighting is a critical factor that directly influences the quality of text vector representation. Feature weighting methods based on category information inadequately capture the relationship between features and categories; specifically, they cannot compare the discriminative power of features with identical category frequencies. Therefore, the intra-class distribution of features must be considered. By incorporating inverse category frequency (ICF) and intra-class entropy into the feature weighting scheme, two supervised feature weighting schemes are constructed. Experimental results on Uyghur text classification corpora demonstrate that the proposed method significantly improves the spatial distribution of samples and enhances the micro-averaged F1 value for Uyghur text classification.

Full Text

Preamble

Feature Weighting Scheme Based on Category Information and Term Entropy

Alimjan Aysa a,b, Yin Xiaoyu b, Kurban Ubul b, Li Zhe a

(a. Network & Information Technology Center; b. School of Information Science & Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: Text vectorization is the foundation of text classification, and feature weighting is one of the crucial factors that directly affects the quality of text vector representation. Existing feature weighting methods based on category information cannot accurately express the relationship between features and categories—specifically, they cannot compare the discriminative power of

features with identical category frequencies. Therefore, the distribution of features within each class must be considered. This paper incorporates inverse category frequency (ICF) and intra-class entropy into the term weighting calculation scheme, constructing two supervised feature weighting schemes. Experimental results on Uyghur text classification corpora demonstrate that the proposed method significantly improves the spatial distribution of samples and enhances the micro-averaged F1 value for Uyghur text classification.

Keywords: text classification; text feature; term weighting; category frequency

0 Introduction

Text classification requires converting natural language texts into an internal representation that computers can process before classifiers can "understand" the content and perform classification. This process is called text vectorization or text representation. Currently, the Vector Space Model (VSM) remains the mainstream approach for text representation. In VSM, a document is represented as a vector $\mathbf{d} = (w_1, w_2, ..., w_n)$ in feature space, where n is the size of the feature set and feature weight w_i indicates the importance of feature i in document d.

Feature weighting schemes in text classification originate from information retrieval (IR). The renowned tf.idf (term frequency and inverse document frequency) algorithm has achieved great success in IR. Because of this success, researchers have applied tf.idf directly to text classification tasks, often using it as the default weighting scheme. Various improved variants have also been proposed. Debole and Sebastiani first proposed a method for constructing supervised feature weighting schemes for text classification by replacing the idf component in tf.idf with feature selection functions such as information gain (IG) and gain ratio (GR). Other researchers combined tf.idf with IG to improve the scheme. Lan et al. proposed tf.rf (term frequency and relevance frequency), which demonstrated improved performance for English text classification. However, tf.rf only considers relevant documents while ignoring the distribution of features in non-relevant documents. Despite this limitation, experiments on standard English corpora showed that tf.rf outperformed other supervised schemes (tf.logOR, tf.χ^2, tf.ig) and traditional methods (tf.idf, tf). Additional supervised weighting schemes include the prob method for imbalanced datasets and icf-based methods for question classification. Recent work has proposed methods based on inverse class space density frequency (ICSDF) that outperform traditional approaches. However, most icf-based algorithms focus only on whether a feature appears in a category without considering its distribution across documents within that category, thereby exaggerating the role of low-frequency documents.

This paper investigates feature weighting for Uyghur text classification. To address the limitations of existing icf-based methods, we incorporate inverse

category frequency (ICF) and intra-class entropy into feature weighting calculations, constructing two supervised feature weighting schemes. We evaluate these schemes on Uyghur text classification datasets and analyze the experimental results.

2 Feature Weighting Schemes Incorporating Category Information and Entropy

We introduce inverse category frequency (ICF) and term entropy into text classification feature weighting.

Category Frequency (CF): CF refers to the number of categories in which feature t_i appears.

Inverse Category Frequency (ICF): ICF is calculated similarly to IDF:

$$icf(t_i) = \log \frac{|C|}{cf(t_i)}$$

where $|C|$ is the total number of categories in the training set. The ICF hypothesis assumes that features appearing in fewer categories carry more category-specific information, emphasizing medium and low-frequency features at the category level while suppressing high-frequency features. However, ICF only considers inter-category distribution without accounting for intra-category distribution. For example, if a feature appears in only a few documents within a category, it cannot well represent that category and should receive lower weight. Therefore, we must also consider the distribution of features within each class.

If a feature term is uniformly distributed within a class, it better represents that class and possesses stronger category discriminative power, warranting higher weight. Conversely, if a feature term appears concentrated in only a few documents within a class, it poorly represents the category and should receive lower weight. Analysis shows that the magnitude of intra-class term entropy correlates with the amount of category information provided by the term—higher entropy indicates more category information and better class representation.

Term Entropy in Class c_k: The entropy $te(t_i, c_k)$ is defined as:

$$te(t_i, c_k) = -\sum_{j=1}^{|c_k|} p(t_i, d_j) \log p(t_i, d_j)$$

where $p(t_i, d_j) = \frac{tf(t_i, d_j)}{\sum_{l=1}^{|c_k|} tf(t_i, d_l)}$ represents the normalized frequency of feature t_i in document d_j of class c_k, $|c_k|$ is the number of documents in class c_k, and $\sum_{j=1}^{|c_k|} tf(t_i, d_j)$ is the total frequency of feature t_i in class c_k.

When feature t_i appears in every document of class c_k, $te(t_i, c_k)$ reaches its maximum value, indicating strongest discriminative power. When t_i appears in only

one document, entropy reaches its minimum, indicating weakest discriminative power. Thus, $te(t_i, c_k)$ effectively reflects the intra-class distribution of features and correlates positively with category discriminative ability.

Based on this analysis, we propose two new feature weighting schemes by incorporating ICF and TE factors:

a) tf.icf.te Scheme:

$$w(t_i, d_j) = tf(t_i, d_j) \times icf(t_i) \times te(t_i, c_k)$$

The tf.icf.te scheme is a hybrid model. While both factors in tf.idf operate at the document level, tf.icf.te calculates TF at the document level, ICF at the category level, and TE measures intra-class distribution.

b) tf.rf.icf.te Scheme:

$$w(t_i, d_j) = tf(t_i, d_j) \times [\alpha \times rf(t_i) + \beta \times icf(t_i) \times te(t_i, c_k)]$$

where $rf(t_i) = \log\left(2 + \frac{a}{\max(1,c)}\right)$, with a being the number of documents containing feature t_i in the positive class and c being the number in the negative class. The tf.rf.icf.te scheme combines four factors: TF represents raw frequency, RF measures distribution between positive and negative classes, ICF measures distribution across categories, and TE measures intra-class distribution.

3 Experiments

3.1 Dataset

We use the Uyghur text dataset Ucorp_A, a balanced dataset containing 10 categories (politics, economics, sports, tourism, education, arts, law, agriculture, medicine, and computer science). Each category includes 300 documents, with 2/3 used for training and 1/3 for testing.

3.2 Evaluation Metrics

Common classification performance metrics include precision, recall, and F1 value:

$$P(\text{precision}) = \frac{\text{correctly classified documents}}{\text{total classified documents}}$$

$$R(\text{recall}) = \frac{\text{correctly classified documents}}{\text{total relevant documents}}$$

$$F1 = \frac{2PR}{P + R}$$

We use micro-averaged F1 value (MicroF1) as the evaluation standard.

3.3 Experimental Results

We first compare the sample space distributions produced by tf.idf, tf.rf, tf.icf.te, and tf.rf.icf.te weighting schemes. We select four feature terms from the Uyghur dataset: “ ”(field), “ ”(competition), “ ”(dawaz), and “ ”(muqam). The first two relate to the “sports” category, while the latter two relate to the “arts” category. and show the weight calculations for these features in different categories.

In the tables, numbers in parentheses after each feature term represent its CF value. tf.idf and tf.icf produce identical weights across categories because they ignore positive/negative class relationships and only consider global characteristics (IDF and ICF). In contrast, tf.rf, tf.icf.te, and tf.rf.icf.te correctly distinguish the four features across categories. Notably, “ ” has CF = 1 (appearing only in the “arts” category). Using tf.icf.te and tf.rf.icf.te, its weight in the “arts” category increases dramatically from 0.023 and 0.044 to 0.521 and 0.667, respectively.

Experiment 1: Sample Space Distribution Comparison

Different weighting methods affect sample distribution in feature space. To evaluate whether our proposed schemes improve sample space distribution, we compare tf.idf, tf.rf, tf.icf.te, and tf.rf.icf.te on dataset Ucorp_A:

a) Preprocess the dataset: tokenization, stopword removal, filtering non-Uyghur characters, removing words shorter than 3 or longer than 24 characters, and filtering words with TF < 3. Apply stemming to remaining words to form the original feature set.

b) Apply the CDDTE (class distribution difference and term entropy) feature selection method to obtain the optimal feature subset.

c) Compute feature weights using tf.idf, tf.rf, tf.icf.te, and tf.rf.icf.te to create four different vector representations.

d) Map the high-dimensional feature space to a lower dimension for visualization and distribution analysis.

[Figure 1: see original paper] shows the sample space distributions on Ucorp_A using the four weighting schemes. All methods improve sample distribution by making intra-class samples more compact and inter-class samples more separated, thereby simplifying the mapping from samples to categories. The tf.rf, tf.icf.te, and tf.rf.icf.te schemes outperform tf.idf, with tf.rf.icf.te showing the best performance.

Experiment 2: Classification Performance Comparison

Feature weighting significantly impacts classification effectiveness. We evaluate the four schemes using Naive Bayes, kNN, Centroid, and SVM classifiers on Ucorp_A.

Using CDDTE for feature selection, we first determined that 2000 features yield optimal MicroF1 values across all classifiers with tf.idf weighting. In this experiment, we use these 2000 features and apply tf.rf, tf.icf.te, and tf.rf.icf.te weighting schemes for classification. presents the results.

The results show that tf.rf, tf.icf.te, and tf.rf.icf.te consistently outperform tf.idf across all four classifiers. tf.rf.icf.te achieves the best performance, with MicroF1 improvements of 6.23% for kNN, 5.08% for Centroid, 2.9% for SVM, and 1.44% for Naive Bayes compared to tf.idf. The tf.rf scheme performs better than tf.idf and tf.icf.te but worse than tf.rf.icf.te. For kNN and Centroid, tf.icf.te improves MicroF1 by 2.53% and 2.11% respectively, while improvements for NB and SVM are minimal due to SVM' s sensitivity to kernel functions rather than feature weights.

These experiments demonstrate that our proposed tf.icf.te and tf.rf.icf.te schemes improve Uyghur text classification performance.

4 Conclusion

This paper proposes two feature weighting schemes, tf.icf.te and tf.rf.icf.te, that incorporate category information and term entropy. Experiments on the Uyghur text dataset Ucorp_A compare these schemes with tf.idf and tf.rf. Visualization of weighted sample spaces qualitatively shows that tf.icf.te and tf.rf.icf.te produce more compact intra-class and more separated inter-class distributions. Classification experiments using Naive Bayes, kNN, Centroid, and SVM demonstrate that the proposed methods improve micro-averaged F1 values for Uyghur text classification.

References

[1] Zhang Aihua, Jing Hongfang, Wang Bin, et al. Research on effects of term weighting factors for text categorization [J]. Journal of Chinese Information Processing, 2010, 24(3): 97-104.

[2] Feng Guozhong, Li Shaoting, Sun Tieli, et al. A probabilistic model derived term weighting scheme for text classification [J]. Pattern Recognition Letters, 2018, 110(1): 23-29.

[3] Chen Kewen, Zhang Zuping, Long Jun, Zhang Hao. Turning from TF-IDF to TF-IGM for term weighting in text classification [J]. Expert Systems With Applications, 2016, 66(1): 245-260.

[4] Chen Kewen, Zhang Zuping, Long Jun. Research on entropy-based term weighting methods in text categorization [J]. Journal of Frontiers of Computer Science & Technology, 2016, 10(9): 1299-1309.

[5] Fattah M A, Sohrab M G. Combined term weighting scheme using FFNN, GA, MR, Sum, & average for text classification [J]. International Journal of Scientific & Engineering Research, 2016, 7(8): 2031-2039.

[6] Debole F, Sebastiani F. Supervised term weighting for automated text categorization [C]// Proc of ACM Symposium on Applied Computing. New York: ACM Press, 2003: 784-788.

[7] Pei Zhuli, Shi Xiaohu, Marchese M, et al. An enhanced text categorization method based on improved text frequency approach and mutual information algorithm [J]. Progress in Natural Science, 2007, 17(12): 1494-1500.

[8] Lan Man, Tan C L, Su Jian, et al. Supervised and traditional term weighting methods for automatic text categorization [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721-735.

[9] Liu Ying, Han Tongloh, Sun Aixin, et al. Imbalanced text classification: a term weighting approach [J]. Expert Systems with Applications, 2009, 36(1): 690-701.

[10] Quan Xiaojun, Liu Wenyin, Qiu Bite et al. Term weighting schemes for question categorization [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(5): 1009-1021.

[11] Jia Longjia, Sun Tieli, Yang Fengqin et al. Class space density based weighting scheme for automated text categorization [J]. Journal of Jilin University, 2017, 35(1): 92-97.

[12] Zhang Ling, Lu Yuliang, Yang Guozheng, et al. Categories-related term weighting method based on term frequency [J]. Application Research of Computers, 2017, 34(2): 386-391.

[13] Li Xueming, Li Hairui, Xue Liang, et al. TFIDF algorithm based on information gain and information entropy [J]. Computer Engineering, 2012, 38(8): 37-40.

[14] Alimjan Aysa, Turgun Ibrahim, Kurban Ubul, et al. Uyghur text feature selection based on class distribution difference and term entropy [J]. Application Research of Computers, 2013, 30(10): 2958-2961.

Note: Figure translations are in progress. See original paper for figures.