From CopeOpi Scores to CopeOpi Vectors: Word Vectors for Multiclass Text Classification

Pei-Shan Tsai
Natural Computing Laboratory
Institute of Computer Science and Engineering
National Chiao Tung University

Abstract

In the era of technology, millions of digital texts are generated every day. To derive useful information from these textual data, text mining has become a popular area of both research and business. One of the most important task of text mining is text classification.

In this thesis, we propose a vector space model for multiclass text classification, the word vectors—CopeOpi vectors. We expand CopeOpi scores which are used in Chinese sentiment analysis, to CopeOpi vectors which can be used in multiclass text classification without the language limit. We verify the functionality of CopeOpi vectors by a series of text classification problems, including sentiment analysis and topic categorization, in both English and Chinese. We make comparisons with several commonly-used features for text classification, and examine these features on different types of machine learning algorithms. The results show that CopeOpi vectors can produce comparable results with a smaller vector size and shorter training time. CopeOpi vectors are effective and efficient features for multiclass text classification.

Keywords: Text classification; Vector space model; Word vector

Computation Scheme 1: CopeOpi Vectors (One-versus-Rest)

Given n corpora of labeled documents $\mathbb{D} = \{\mathbb{D}_{c_1}, \mathbb{D}_{c_2}, \ldots, \mathbb{D}_{c_n}\}$, and the corresponding classes $\mathbb{C} = \{c_1, c_2, \ldots, c_n\}$.

- $\mathbb{D}_{c_i} = \{\langle d, c\rangle \mid d \text{ is a document labeled as class } c = c_i\}$
  - the vocabulary $\mathbb{V}_{c_i}$ is a set of unique words in $\mathbb{D}_{c_i}$

For each word $w_i \in \cup_{c\in\mathbb{C}}\mathbb{V}_c$, we can compute its CopeOpi vector $\overrightarrow{\mathcal{COP}}_{w_i}$ by one-versus-rest strategy.

For each class $c_j \in \mathbb{C}$, we can construct two opposite sets,

- the positive set $\mathbb{P}_{w_i}^{c_j} = \{c_j\}$
  - the positive corpus $\mathbb{D}_{\mathbb{P}w_i}^{c_j} = \{\mathbb{D}_{c_j}\}$
  - the positive vocabulary $\mathbb{V}_{\mathbb{P}w_i}^{c_j} = \mathbb{V}_{c_j}$
- the negative set $\mathbb{N}_{w_i}^{c_j} = \mathbb{C} \setminus \{c_j\}$
  - the negative corpus $\mathbb{D}_{\mathbb{N}w_i}^{c_j} = \mathbb{D} \setminus \{\mathbb{D}_{c_j}\}$
  - the negative vocabulary $\mathbb{V}_{\mathbb{N}w_i}^{c_j} = \cup_{c\in\mathbb{N}_{w_i}^{c_j}} \mathbb{V}_c$

and compute the augmented CopeOpi score $\mathcal{COP}_{w_i}^{c_j}$ of word w_i with respect to class c_j based on these two opposite sets.

$$\mathcal{P}_{w_i}^{c_j} = \frac{fp_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j}} fp_w^{c_j}}{fp_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j}} fp_w^{c_j} + fn_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j}} fn_w^{c_j}}$$

$$\mathcal{N}_{w_i}^{c_j} = \frac{fn_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j}} fn_w^{c_j}}{fp_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j}} fp_w^{c_j} + fn_{w_i}^{c_j} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j}} fn_w^{c_j}}$$

$$\mathcal{COP}_{w_i}^{c_j} = \mathcal{P}_{w_i}^{c_j} - \mathcal{N}_{w_i}^{c_j}$$

where $\mathcal{P}_{w_i}^{c_j}$ and $\mathcal{N}_{w_i}^{c_j}$ are the normalized probabilities of word w_i being in class $\mathbb{P}_{w_i}^{c_j}$ and being in class $\mathbb{N}_{w_i}^{c_j}$; $fp_{w_i}^{c_j}$ and $fn_{w_i}^{c_j}$ are the frequencies of word w_i in corpus $\mathbb{D}_{\mathbb{P}w_i}^{c_j}$ and in corpus $\mathbb{D}_{\mathbb{N}w_i}^{c_j}$; the CopeOpi score $\mathcal{COP}_{w_i}^{c_j}$ of word w_i with respect to class c_j is defined as the difference of the two opposite normalized probabilities, ranged from $+1$(being in class $\mathbb{P}_{w_i}^{c_j}$) to -1(being in class $\mathbb{N}_{w_i}^{c_j}$).

The CopeOpi vector $\overrightarrow{\mathcal{COP}}_{w_i}$ of word w_i is composed of these n augmented CopeOpi scores.

$$\overrightarrow{\mathcal{COP}}_{w_i} = (\mathcal{COP}_{w_i}^{c_1}, \mathcal{COP}_{w_i}^{c_2}, \ldots, \mathcal{COP}_{w_i}^{c_n})$$

Computation Scheme 2: CopeOpi Vectors (One-versus-One)

Given n corpora of labeled documents $\mathbb{D} = \{\mathbb{D}_{c_1}, \mathbb{D}_{c_2}, \ldots, \mathbb{D}_{c_n}\}$, and the corresponding classes $\mathbb{C} = \{c_1, c_2, \ldots, c_n\}$.

- $\mathbb{D}_{c_i} = \{\langle d, c\rangle \mid d \text{ is a document labeled as class } c = c_i\}$
  - the vocabulary $\mathbb{V}_{c_i}$ is a set of unique words in $\mathbb{D}_{c_i}$

For each word $w_i \in \cup_{c\in\mathbb{C}}\mathbb{V}_c$, we can compute its CopeOpi vector $\overrightarrow{\mathcal{COP}}_{w_i}$ by one-versus-one strategy.

For each class-pair $c_j, c_k \in \mathbb{C}$, $1 \le j < k \le n$, we can construct two opposite sets,

- the positive set $\mathbb{P}_{w_i}^{c_j,c_k} = \{c_j\}$
  - the positive corpus $\mathbb{D}_{\mathbb{P}w_i}^{c_j,c_k} = \{\mathbb{D}_{c_j}\}$
  - the positive vocabulary $\mathbb{V}_{\mathbb{P}w_i}^{c_j,c_k} = \mathbb{V}_{c_j}$
- the negative set $\mathbb{N}_{w_i}^{c_j,c_k} = \{c_k\}$
  - the negative corpus $\mathbb{D}_{\mathbb{N}w_i}^{c_j,c_k} = \{\mathbb{D}_{c_k}\}$
  - the negative vocabulary $\mathbb{V}_{\mathbb{N}w_i}^{c_j,c_k} = \mathbb{V}_{c_k}$

and compute the augmented CopeOpi score $\mathcal{COP}_{w_i}^{c_j,c_k}$ of word w_i with respect to class-pair c_j, c_k based on these two opposite sets.

$$\mathcal{P}_{w_i}^{c_j,c_k} = \frac{fp_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j,c_k}} fp_w^{c_j,c_k}}{fp_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j,c_k}} fp_w^{c_j,c_k} + fn_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j,c_k}} fn_w^{c_j,c_k}}$$

$$\mathcal{N}_{w_i}^{c_j,c_k} = \frac{fn_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j,c_k}} fn_w^{c_j,c_k}}{fp_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{P}w_i}^{c_j,c_k}} fp_w^{c_j,c_k} + fn_{w_i}^{c_j,c_k} / \sum_{w\in\mathbb{V}_{\mathbb{N}w_i}^{c_j,c_k}} fn_w^{c_j,c_k}}$$

$$\mathcal{COP}_{w_i}^{c_j,c_k} = \mathcal{P}_{w_i}^{c_j,c_k} - \mathcal{N}_{w_i}^{c_j,c_k}$$

where $\mathcal{P}_{w_i}^{c_j,c_k}$ and $\mathcal{N}_{w_i}^{c_j,c_k}$ are the normalized probabilities of word w_i being in class $\mathbb{P}_{w_i}^{c_j,c_k}$ and being in class $\mathbb{N}_{w_i}^{c_j,c_k}$; $fp_{w_i}^{c_j,c_k}$ and $fn_{w_i}^{c_j,c_k}$ are the frequencies of word w_i in corpus $\mathbb{D}_{\mathbb{P}w_i}^{c_j,c_k}$ and in corpus $\mathbb{D}_{\mathbb{N}w_i}^{c_j,c_k}$; the CopeOpi score $\mathcal{COP}_{w_i}^{c_j,c_k}$ of word w_i with respect to class-pair c_j, c_k is defined as the difference of the two opposite normalized probabilities, ranged from $+1$(being in class $\mathbb{P}_{w_i}^{c_j,c_k}$) to -1(being in class $\mathbb{N}_{w_i}^{c_j,c_k}$).

The CopeOpi vector $\overrightarrow{\mathcal{COP}}_{w_i}$ of word w_i is composed of these $\frac{1}{2}n(n-1)$ augmented CopeOpi scores.

$$\overrightarrow{\mathcal{COP}}_{w_i} = (\mathcal{COP}_{w_i}^{c_1,c_2}, \mathcal{COP}_{w_i}^{c_1,c_3}, \ldots, \mathcal{COP}_{w_i}^{c_{n-1},c_n})$$