Postprint of Text Sentiment Analysis Based on Hybrid Mutual Information Algorithm

Authors: Wang Yi, Dai Yue-Ming

Date: 2018-12-13T00:00:00+00:00

Abstract

To address the phenomenon of positive and negative correlations in mutual information (MI) feature selection methods and the problem of not considering term frequency of feature items within different categories, a hybrid mutual information (HMI) feature selection algorithm is proposed. This algorithm introduces an inverse document frequency coefficient and an inter-class term frequency information coefficient, enabling effective utilization of term frequency information across the entire document as well as among classes; it also introduces a positive-negative correlation coefficient to distinguish between positive and negative correlations and effectively utilize them. Experimental comparisons demonstrate that the hybrid mutual information algorithm can effectively improve the quality of feature selection, thereby enhancing the effectiveness of text sentiment analysis.

Full Text

Preamble

Vol. 37 No. 2
Application Research of Computers
Accepted Paper

Text Sentiment Analysis Based on Hybrid Mutual Information Algorithm

WANG Yi, DAI Yueming
(School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China)

Abstract: The mutual information (MI) feature selection method suffers from the phenomenon of positive and negative correlation and fails to consider word

frequency across different categories. To address these issues, this paper proposes a hybrid mutual information feature selection algorithm (HMI). This algorithm introduces an inverse document frequency coefficient and an inter-class word frequency information coefficient, enabling effective utilization of word frequency information both globally across the entire document collection and locally between classes. Additionally, a positive and negative correlation coefficient is introduced to distinguish and effectively leverage both types of correlations. Experimental comparisons demonstrate that the hybrid mutual information algorithm can effectively improve the quality of feature selection and thereby enhance the effectiveness of text sentiment analysis.

Keywords: mutual information; feature selection; positive and negative correlation; word frequency information; sentiment analysis

0 Introduction

With the continuous development of technology and the increasing popularity of the Internet, the demand for data analysis has grown substantially, accompanied by a surge in reviews about products and services. Consequently, sentiment analysis of these reviews and evaluations has become a hot research topic [1]. Text sentiment analysis examines the emotional orientation of texts and mines effective information such as authors' viewpoints and attitudes [2], which is why it is also known as opinion mining. Currently, the two main approaches to text sentiment analysis are sentiment dictionary-based methods and machine learning-based methods, with the latter being the mainstream approach for sentiment classification [3]. In text sentiment analysis, textual data is typically represented using the Vector Space Model (VSM) [4], which transforms unstructured text data into structured data that computers can process. In typical datasets, the number of feature terms can reach tens of thousands, and in larger datasets, even millions. Therefore, a key challenge in text sentiment analysis is how to reduce the dimensionality of the feature space while improving the effectiveness of text sentiment classification [5]. Feature selection has naturally become a crucial component of text sentiment analysis.

The primary goal of feature selection is to improve classification accuracy by removing noise [6] and selecting high-quality, representative terms. Common feature selection methods include Chi-square statistic (CHI), document frequency (DF), information gain (IG), and mutual information (MI). Among these, MI has become an important feature selection method due to its low time complexity, ease of understanding, and convenient implementation [7]. However, traditional MI algorithms produce lower-quality feature terms because they ignore word frequency factors [8]. Moreover, when calculating the mutual information value of feature terms, they overlook negatively correlated features, causing the values of negatively correlated features to significantly weaken the overall feature value and thereby reducing the accuracy of the MI algorithm.

In summary, to address the shortcomings of the MI algorithm in feature selection, this paper introduces an inverse document frequency coefficient, an inter-class frequency coefficient, and a positive/negative correlation coefficient, proposing a hybrid mutual information (HMI) feature selection method. Both theoretical analysis and experimental results demonstrate that this algorithm can effectively utilize word frequency information and positive/negative correlation information to improve the quality of feature selection and thus enhance the accuracy of sentiment classification.

---

1 Mutual Information Feature Selection Method

Mutual information (MI) is a statistical algorithm commonly used to measure the degree of association between two statistical variables [9]. In text sentiment analysis, the MI algorithm is typically employed to calculate the degree of association between feature terms in the text and various categories. When the association between a feature term and a category is stronger, their mutual information value is larger, indicating that the feature term is more representative of that category [9]. Let t_k denote the set of feature terms, where $k = 1, 2, \ldots, m$, and let c_j denote the set of categories in the training set, where $j = 1, 2, \ldots, r$. The mutual information value between t_k and c_j is calculated as follows:

$$\mathrm{MI}(t_k, c_j) = \log \frac{p(t_k, c_j)}{p(t_k)p(c_j)} = \log \frac{p(t_k|c_j)}{p(t_k)}$$

where $p(t_k, c_j)$ represents the probability of texts containing feature t_k and belonging to category c_j in the training set, $p(t_k)$ represents the probability of texts containing feature t_k in the entire training set, $p(c_j)$ represents the probability of texts belonging to category c_j in the training set, and $p(t_k|c_j)$ represents the probability of texts containing feature t_k within category c_j. For training sets with multiple categories, the mutual information between feature term t_k and all categories in the training set is calculated as:

$$\mathrm{MI}(t_k) = \sum_{j=1}^{r} p(c_j) \cdot \mathrm{MI}(t_k, c_j) = \sum_{j=1}^{r} p(t_k, c_j) \log \frac{p(t_k|c_j)}{p(t_k)}$$

The MI algorithm calculates the ratio of the number of texts containing feature term t_k to the number of texts in each category c_j in the training set. Its most important characteristic is that it considers the co-occurrence frequency between different feature terms and a specific category, effectively utilizing text category information [10]. However, the MI method also has some obvious shortcomings. For instance, in Equation (2), the differences in frequency counts of each feature term across different categories are not reflected, nor are the relationships between frequency counts of texts containing the feature term in the training set

considered. In text sentiment analysis, the correlation between feature terms and categories can be divided into positive correlation and negative correlation. Positively correlated feature terms play a primary role in text sentiment classification [11], but negatively correlated features also have an important impact on the final classification results. As shown in Equation (2), the mutual information values of positive and negative correlations cancel each other out, thereby ignoring the role of negative correlations.

---

2 Hybrid Mutual Information Algorithm

Based on the above analysis, this section introduces three coefficients to address the limitations of traditional MI: an inter-class word frequency information coefficient (α), an inverse document frequency coefficient (β), and a positive/negative correlation coefficient (γ). The final hybrid mutual information (HMI) formula integrates all three coefficients.

2.1 Inter-Class Word Frequency Information Coefficient

In traditional MI feature selection, only the frequency of feature terms within a class is considered [12]. However, the role of term frequency in text sentiment analysis is not only reflected within classes but also plays a crucial role between classes. If a feature term is more representative of a particular class, it should be concentrated in that class—that is, its term frequency should be relatively high in that class and appear as little as possible in other classes.

Assume feature term t_k is a feature of category c_j. Then t_k should appear as frequently as possible in category c_j and as rarely as possible in other categories c_q (where $q \neq j$). In theoretical terms, for feature terms with strong class representational ability, the standard deviation across different categories should be as large as possible. Based on this consideration, this paper introduces the inter-class word frequency information coefficient α on the basis of Equation (2). The definition of α is as follows:

$$\alpha = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left[\frac{tf_j(t_k)}{\sum_{i=1}^{m} tf_i(t_k)} - \frac{1}{m}\right]^2}$$

where $tf_j(t_k)$ represents the frequency of feature t_k in category j, and m represents the total number of categories.

2.2 Inverse Document Frequency Coefficient

The introduction of the inter-class word frequency information coefficient α reveals that when a feature term is concentrated in the texts of a particular class, it has strong representational power for that class. However, traditional

MI methods also ignore the document frequency of feature terms. For example, words like "he," "you," and "is" may appear in many texts, but such words have low discriminative power for text classification [13]. Therefore, if a feature term appears in most texts, it means the feature term has weaker ability to distinguish text categories. To address this issue and increase the discriminative power of feature terms, this paper introduces the inverse document coefficient β.

The definition of β is as follows:

$$\beta = \log \frac{N}{f(t_k) + 0.01}$$

where N represents the total number of documents in the training set, and $f(t_k)$ represents the number of texts containing feature term t_k. The addition of 0.01 in the denominator ensures it never becomes zero, guaranteeing the coefficient' s validity. In this formula, since $N \geq f(t_k)$ always holds, when more texts contain feature term t_k (i.e., $f(t_k)$ is larger), $f(t_k)$ approaches N, and coefficient β approaches 0, meaning its impact on the MI value of that feature term diminishes. By introducing the inverse document frequency coefficient, the influence of common words as feature terms on the final classification result is reduced, thereby improving the effectiveness of feature selection.

2.3 Positive/Negative Correlation Coefficient

As shown in Equation (1), when calculating the mutual information value between feature term t_k and category c_j, if $p(t_k|c_j) > p(t_k)$, then $\mathrm{MI}(t_k, c_j) > 0$, indicating that feature term t_k and category c_j are positively correlated. This means that as $p(t_k|c_j)$ increases and $p(t_k)$ decreases, the ability of feature term t_k to represent category c_j becomes stronger. Conversely, if $p(t_k|c_j) < p(t_k)$, then $\mathrm{MI}(t_k, c_j) < 0$, indicating negative correlation between feature term t_k and category c_j. This means that as $p(t_k|c_j)$ decreases and $p(t_k)$ increases, the amount of information between feature term t_k and category c_j decreases, and the representational ability of t_k for c_j becomes weaker.

As seen in Equation (2), the negatively correlated portion of a feature term' s value weakens its final MI value when calculating the MI value for the category set. In text sentiment classification, positively correlated features help improve final accuracy, while negatively correlated features help improve final recall [14]. Therefore, the role of negatively correlated features cannot be ignored [15]. To address this phenomenon, this paper introduces the positive/negative correlation coefficient γ to adjust the positive/negative correlation issue in the MI method.

In category c_j $(j = 1, 2, \ldots, r)$, we first define:

$$\bar{f}(t_k) = \frac{1}{r} \sum_{j=1}^{r} f_j(t_k)$$

which represents the average number of texts containing feature term t_k across all categories, where $f_j(t_k)$ represents the number of texts in category c_j that contain feature term t_k.

When $p(t_k|c_j) > p(t_k)$ (positive correlation), γ is defined as:

$$\gamma = \omega \times \frac{f_j(t_k) - \bar{f}(t_k)}{\bar{f}(t_k)}$$

When $p(t_k|c_j) < p(t_k)$ (negative correlation), γ is defined as:

$$\gamma = (1 - \omega) \times \frac{f_j(t_k) - \bar{f}(t_k)}{\bar{f}(t_k)}$$

where ω is a 调节因子 (regulation factor) with a theoretical range of 0.1~0.9, used to adjust the influence of positively and negatively correlated features, ensuring that both types of features fully contribute to the final sentiment classification. In the coefficient γ, when a feature term is negatively correlated with a category, the term $\frac{f_j(t_k)-\bar{f}(t_k)}{\bar{f}(t_k)}$ effectively handles cases where the feature term appears less frequently in that class.

2.4 Final HMI Formula

In summary, the definition of Hybrid Mutual Information (HMI) is as follows:

$$\mathrm{HMI}(t_k) = \alpha \times \beta \times \sum_{j=1}^{r} \gamma \times p(t_k|c_j) \log \frac{p(t_k|c_j)}{p(t_k)}$$

The HMI algorithm pseudocode is as follows:

```
1. for each document dj  D do
2.   for each word tk  dj do
3.     // Calculate term frequency
4.   end for
5. end for
6. for each category Cj  C do
7.   // Calculate category statistics
8. end for
9. for each document dj  D do
10.  // Calculate document-level features
11. end for
12. for each document dj  Cj do
13.  // Calculate category-level features
14. end for
15. for each document dj  D do
```

```
16.  if word   dj then
17.    // Update word counts
18.  end if
19. end for
20. for each document dj   D do
21.  // Calculate final statistics
22. end for
23.   = sqrt(square(tc - (sum(tk)/Ck))/Ck)
24.   = log[N/(count + 0.01)]
25. if (dk/Dj) >= (count/N) then
26.     =   × [(dk - count/Ck)/(count/Ck)]
27. else
28.     = (1- ) × [(dk - count/Ck)/(count/Ck)]
29. end if
```

3 Experiments

3.1 Experimental Datasets

This study employs two datasets for experiments: the Tan Songbo hotel management review corpus and the Midea air conditioner review corpus. Each corpus contains 4,000 review comments, divided into two categories: positive and negative reviews, with 2,000 positive (pos) reviews and 2,000 negative (neg) reviews. To verify the effectiveness of the algorithm, cross-validation is adopted, with 80% of the corpus data used as the training set and the remaining 20% as the test set. shows examples from the training set used in this study, while shows examples from the test set.

3.2 Evaluation Metrics

In text sentiment analysis, commonly used evaluation metrics include precision (also called accuracy), recall (also called sensitivity), and the F1-score, which combines precision and recall. This paper employs these three metrics to evaluate experimental results. The classification outcomes in text sentiment analysis can be categorized into four situations, as shown in .

lists the actual and predicted results for the test set examples from . In the table, TP refers to the number of texts predicted as positive that are actually positive, as shown by example S4 in ; FP refers to the number of texts predicted as positive but actually negative, as shown by example S6; FN refers to the number of texts predicted as negative but actually positive, as shown by example S5; and TN refers to the number of texts predicted as negative that are actually negative, as shown by example S7.

The definitions of precision and recall are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1-score combines precision and recall:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3 Experimental Procedure

The experimental steps are as follows:

a) **Data Preprocessing**: The corpus is annotated, with positive and negative corpora merged into a single document for segmentation. This study uses the commonly used jieba Chinese word segmentation tool. The segmented results for the training set examples in are shown in .

b) **Stopword Removal**: Stopwords, punctuation marks, and other factors irrelevant to text sentiment classification are removed. The results after stopword removal for the training set examples are shown in .

c) **Feature Selection**: Three feature selection methods are applied: Mutual Information (MI), Hybrid Mutual Information (HMI), and Chi-square statistic (CHI).

d) **Feature Representation**: The Bag-of-Words (BOW) model is used to represent features, and the Vector Space Model (VSM) converts text data into structured data.

e) **Classification**: The Support Vector Machine (SVM) classifier is adopted for training and testing due to its simple structure and global optimality, making it a mainstream classifier in text sentiment analysis. The three feature selection methods are compared and analyzed.

The overall experimental flow is illustrated in [Figure 1: see original paper].

3.4 Results and Analysis

This study conducts comparative experiments using three different feature selection methods: CHI, MI, and the proposed HMI. The BOW model is used for feature representation, and precision, recall, and F1-score are calculated at dimensions of 2000, 3000, 4000, 5000, 6000, 7000, 8000, and 9000. Since positively correlated features play a primary role, the regulation factor ω in Equation (8) is tested with values of 0.5, 0.6, 0.7, 0.8, and 0.9. Through multiple comparative experiments, $\omega = 0.8$ yields the best results.

through show the precision, recall, and F1-score comparisons for the hotel management review dataset across different dimensions. The results indicate that HMI significantly outperforms the other two methods across all three metrics. Specifically, precision improves by 5% over MI and 11% over CHI; recall improves by 9% over both MI and CHI; and F1-score improves by 6% over MI and 9% over CHI. These results demonstrate that HMI effectively enhances text sentiment classification performance.

through show the corresponding comparisons for the Midea air conditioner review dataset. Again, HMI significantly outperforms MI and CHI across all metrics: precision improves by 12% over MI and 11% over CHI; recall improves by 12% over MI and 6% over CHI; and F1-score improves by 8% over MI and 9% over CHI. These results confirm that HMI consistently improves text sentiment classification across different datasets.

To visualize the performance changes across different feature dimensions, [Figure 2: see original paper] through [Figure 4: see original paper] present line charts for the hotel management review dataset, while [Figure 5: see original paper] through [Figure 7: see original paper] present corresponding charts for the Midea air conditioner review dataset.

As shown in [Figure 2: see original paper] through [Figure 4: see original paper], the performance of MI and CHI begins to stabilize at around 7000 dimensions, whereas HMI maintains consistent performance across all dimensions, consistently outperforming the other two algorithms. Similarly, [Figure 5: see original paper] through [Figure 7: see original paper] show that while all three algorithms exhibit upward trends in precision and recall, HMI consistently achieves superior overall performance with better stability in F1-score across dimensions.

In conclusion, the HMI algorithm achieves significantly higher precision, recall, and F1-score compared to MI and CHI methods, while demonstrating strong stability across different feature dimensions. Therefore, the HMI feature selection algorithm can effectively improve feature selection quality and enhance text sentiment classification performance.

---

4 Conclusion

This paper proposes a Hybrid Mutual Information (HMI) feature selection algorithm to address the issues of positive/negative correlation phenomena and the neglect of word frequency information in traditional MI-based feature selection methods. By introducing inverse document frequency, inter-class word frequency, and positive/negative correlation indicators, the algorithm effectively utilizes word frequency information within the MI framework and appropriately leverages the important roles of both positive and negative correlations. Experimental results demonstrate that the HMI method significantly outperforms other feature selection methods and achieves promising results in text sentiment

classification.

References

[1] Cherry C, Mohammad S. Binary classifiers and latent sequence models for emotion detection in suicide notes [J]. Journal of Biomedical Informatics Insights, 2012, 5(S1): 147-154.

[2] Atiyeh M, Hossein M M. Robust feature selection from microarray data based on cooperative game theory and qualitative mutual information [J]. Advances in Bioinformatics, 2016, 2016(1): 1-16.

[3] 李平, 戴月明, 王艳. 基于混合卡方统计量与逻辑回归的文本情感分析 [J]. 计算机工程, 2017, 43(12): 192-196. (Li Ping, Dai Yueming, Wang Yan. Text emotion analysis based on mixed chi-square statistics and logical regression [J]. Computer Engineering, 2017, 43(12): 192-196.)

[4] Tang Jian, Zhou Shuigeng. A new approach for feature selection from microarray data based on information [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2016, 13(6): 1004-1015.

[5] Bidi N, Elberichi Z. Feature selection for text classification using genetic algorithms [C]//Proc of the 8th International Conference on Modelling, Identification and Control. Piscataway, NJ: IEEE Press, 2016.

[6] 朱颢东, 陈宁, 李红婵. 优化的互信息特征选择方法 [J]. 计算机工程与应用, 2010, 46(26): 122-124. (Zhu Yidong, Chen Ning, Li Hongchen. Optimized mutual information feature selection method [J]. Computer Engineering and Application, 2010, 46(26): 122-124.)

[7] 陶永才, 赵国桦, 石磊, 等. 一种改进的 MapReduce 互信息文本特征选择机制 [J]. 小型微型计算机系统, 2018, 39(3): 433-438. (Tao Yongcai, Zhao Guohua, Shi Lei, et al. Improved MapReduce mutual information text feature selection mechanism [J]. Minicomputer System, 2018, 39(3): 433-438.)

[8] Li Kewen, Yu Mingxiao, Liu Lu, et al. Feature selection method based on weighted mutual information for imbalanced Data [J]. International Journal of Software Engineering and Knowledge Engineering, 2018, 28(8): 1177-1194.

[9] Coelho F, Braga A P, Verleysen M. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems [J]. International Journal of Computational Intelligence Systems, 2016, 9(4): 726-733.

[10] 刘海峰, 陈琦, 张以皓. 一种基于互信息的改进文本特征选择 [J]. 计算机工程与应用, 2012, 48(25): 1-4. (Liu Haifeng, Chen Qi, Zhang Yihao. An improved text feature selection based on mutual information [J]. Computer Engineering and Application, 2012, 48(25): 1-4.)

[11] 林少波, 杨丹, 徐玲. 基于类别相关的新文本特征提取方法 [J]. 计算机应用研究, 2012, 29(5): 1680-1683. (Lin Shaobo, Yang Dan, Xu Ling. A new text feature extrac-

tion method based on category correlation [J]. Computer Application Research, 2012, 29(5): 1680-1683.)

[12] Calvo B, Larrariaga P, Lozano J A. Feature subset selection from positive and unlabeled example [J]. Patten Recognition Letters, 2009, 30(11): 1027-1036.

[13] Lin Yaojin, Hu Qinghua, Liu Jinghua, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. IEEE Trans on Fuzzy Systems, 2017, 25(6): 1491-1507.

[14] Bostani H, Sheikhan M. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems [J]. Soft Computing, 2017, 21(9): 2307-2324.

[15] 辛竹, 周亚建. 文本分类中互信息特征选择方法的研究与算法改进 [J]. 计算机应用, 2013, 33(S2): 116-118. (Xin Zhu, Zhou Yajian. Research and improvement of mutual information feature selection method in text classification [J]. Computer applications, 2013, 33(S2): 116-118.)

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