A Feature Selection Method Based on the Combination of Improved CHI and Weighted ECE (Postprint)

Authors: Cai Zhen, Gao Jian, Qin Xiaojun

Date: 2018-07-09T00:00:00+00:00

Abstract

Regarding the Chi-square statistic (CHI) and expected cross entropy (ECE) feature selection methods in text classification, this paper analyzes their characteristics and shortcomings. To address the issue of poor classification performance of traditional CHI and ECE methods on imbalanced datasets, we propose an improved CHI method (pCHI) by introducing adjustment factors and eliminating negative correlation influences, and remedy the defect of ECE' s propensity to select high-frequency features with weak discriminative capability through weighting (ωECE). Based on the synthesis of these two improved methods, we further propose a feature selection method that combines improved CHI and weighted ECE (pCHIωECE). Verified through comparative experiments, the pCHIωECE method achieves superior precision and F1-score compared to CHI, ECE, pCHI, and ωECE methods, and demonstrates enhanced stability in dimensionality reduction.

Full Text

Preamble

Feature Selection Method Based on Combining Improved CHI and Weighted ECE

Cai Zhen, Gao Jian, Qin Xiaojun
(Jiangnan Institute of Computing Technology, Wuxi, Jiangsu 214083, China)

Abstract: This paper analyzes the characteristics and deficiencies of chi-square statistics (CHI) and expected cross-entropy (ECE) methods for feature selection in text classification. To address the poor classification performance of traditional CHI and ECE methods on imbalanced datasets, we present an improved CHI method (pCHI) by introducing adjustment factors and removing negative

correlation influences, and propose a weighted ECE method (ωECE) to compensate for the tendency of ECE to select high-frequency features with weak discriminative ability. By synthesizing these two improved methods, we further propose a feature selection method based on combining improved CHI and weighted ECE (pCHIωECE). Comparative experiments demonstrate that the pCHIωECE method achieves superior precision rates and F1 values compared to CHI, ECE, pCHI, and ωECE methods, while also exhibiting better dimensionality reduction stability.

Keywords: chi-square statistics; expected cross-entropy; feature selection; text classification

1 Related Work

Text classification [1] refers to the technique of categorizing large volumes of text into one or more predefined categories, with widespread applications in data mining, machine learning, information retrieval, and other domains. The process typically involves document representation, feature selection, and classifier training. Feature selection aims to identify an optimal feature subset from the original feature space. Due to the "curse of dimensionality"[2] in text features and the presence of irrelevant features (noise), feature selection is particularly crucial for text classification effectiveness.

Text classification commonly employs the Vector Space Model (VSM) for representation. Let a document be represented as $\langle\langle MATH_1 \rangle\rangle$, where N is the total number of documents. $\langle\langle MATH_2 \rangle\rangle$ denotes the probability of feature t appearing in k different categories, with k being the total number of categories, i.e., $\langle\langle MATH_3 \rangle\rangle$. $\langle\langle MATH_4 \rangle\rangle$ represents the conditional probability that a document belongs to class i given that it contains feature t, and $\langle\langle MATH_5 \rangle\rangle$. $\langle\langle MATH_6 \rangle\rangle$ is the global probability of class i, and F(t) is the number of documents containing feature t.

Common feature selection algorithms for text classification are designed based on information theory and statistical principles, including Gini index, document frequency, information gain, mutual information, chi-square statistics, expected cross-entropy, and linear discriminant analysis. References [3-6] provide detailed discussions of these methods and their characteristics. Addressing the problem that traditional chi-square statistics and expected cross-entropy methods yield poor classification results on imbalanced datasets and under noise interference, this paper proposes a feature selection method based on combining improved CHI and weighted ECE. Comparative experiments demonstrate that this method effectively improves text classification accuracy.

1.1 Chi-Square Statistics (CHI)

Chi-square statistics [7] measure the independence between feature t and a specific category i. The chi-square statistic between feature t and category i is defined as:

$$\langle\langle MATH_7 \rangle\rangle$$

where $\langle\langle MATH_8 \rangle\rangle$ gives the case where a document belongs to class i and contains feature t, while $\langle\langle MATH_9 \rangle\rangle$ gives the case where these conditions are not simultaneously satisfied. In fact, whether $\langle\langle MATH_{10} \rangle\rangle$ is large or small depends on the degree of correlation between category i and feature t.

The global chi-square statistic for feature t can be calculated using either a weighted average or maximum value, with the formulas:

$$\langle\langle MATH_{11} \rangle\rangle$$

$$\langle\langle MATH_{12} \rangle\rangle$$

The chi-square statistic is a standardized value that provides strong discriminative power among features within the same category. If feature t and category i are independent, then $\langle\langle MATH_{13} \rangle\rangle$. The stronger the correlation between feature t and category i, the larger the value of $\langle\langle MATH_{14} \rangle\rangle$, indicating that feature t contains more information relevant to category i.

Features and categories exhibit both positive and negative correlations. Let $\langle\langle MATH_{15} \rangle\rangle$. If $\langle\langle MATH_{16} \rangle\rangle$, feature t is positively correlated with category i, and larger values of $\langle\langle MATH_{17} \rangle\rangle$ indicate a higher probability that documents containing feature t belong to class i. Conversely, if $\langle\langle MATH_{18} \rangle\rangle$, feature t is negatively correlated with category i, and larger values of $\langle\langle MATH_{19} \rangle\rangle$ indicate a higher probability that documents containing feature t do not belong to class i.

Traditional CHI statistics only consider the number of documents containing feature words across the entire document collection, without accounting for the frequency of feature words within individual documents, thereby exaggerating the role of low-frequency words. Classification performance degrades significantly on imbalanced samples.

1.2 Information Gain (IG) and Expected Cross-Entropy (ECE)

Information gain [1] calculates the information gain of feature t for category i by counting document frequencies where feature t appears or does not appear in the category. It considers the difference in information entropy before and after feature t appears. The information gain formula for feature t is:

$$\langle\langle MATH_{20}\rangle\rangle$$

where $\langle\langle MATH_{21}\rangle\rangle$ and $\langle\langle MATH_{22}\rangle\rangle$. A larger IG(t) value indicates greater discriminative ability of feature t.

Expected cross-entropy [8] is similar to information gain, but differs in that ECE only calculates features that appear in the text, without considering cases where features are absent. The expected cross-entropy for feature t is:

$$\langle\langle MATH_{23}\rangle\rangle$$

where $\langle\langle MATH_{24}\rangle\rangle$ represents the frequency of feature t in category i, $\langle\langle MATH_{25}\rangle\rangle$ represents the frequency of feature t across all categories, and $\langle\langle MATH_{26}\rangle\rangle$ represents the ratio of feature t' s frequency in a specific category i to its frequency in all other categories. Adding 1 to the denominator in equation (6) prevents division by zero when feature t appears only in category i. If $\langle\langle MATH_{27}\rangle\rangle$ is larger, feature t appears more frequently in category i and less frequently in other categories, indicating that the feature provides greater discriminative ability for the category. Conversely, smaller values indicate weaker discriminative ability.

Similarly, a larger expected cross-entropy ECE(t) value indicates stronger discriminative ability of feature t. References [9,10] show that while the absence of feature t may contribute to category determination, this contribution is often far outweighed by the interference it introduces. Particularly under highly imbalanced category and feature distributions, if $\langle\langle MATH_{28}\rangle\rangle$—meaning the vast majority of features are absent—the value of IG(t) in equation (1) is determined by $\langle\langle MATH_{29}\rangle\rangle$, causing IG(t) to favor words with low frequency. ECE' s exclusion of the impact of feature absence in same-class documents is precisely why it outperforms IG.

---

2 Proposed Methods

2.1 CHI Analysis and Improvement

Equation (1) shows that the CHI method uses $\langle\langle MATH_{30}\rangle\rangle$ to account for cases where a document belongs to class i and contains feature t, and $\langle\langle MATH_{31}\rangle\rangle$ for cases where both conditions are simultaneously satisfied or unsatisfied. To address CHI' s limitation of exaggerating low-frequency words, this paper introduces a feature frequency adjustment factor α to reduce interference from low-frequency features in text classification. The calculation formula for α is:

$$\langle\langle MATH_{32}\rangle\rangle$$

where $\langle\langle MATH_{33}\rangle\rangle$ represents the frequency of feature t in category i, and $\langle\langle MATH_{34}\rangle\rangle$ represents the total frequency of feature t across all categories. The factor $\langle\langle MATH_{35}\rangle\rangle$ represents the ratio of feature t' s frequency in a specific category i to its frequency in all other categories. A larger $\langle\langle MATH_{36}\rangle\rangle$ indicates that feature t appears more frequently in category i and less frequently in other categories, allowing the feature to provide greater discriminative ability for the category. Conversely, smaller $\langle\langle MATH_{37}\rangle\rangle$ values indicate weaker discriminative ability of feature t.

By further removing negative correlation cases between features and categories and combining equations (1) and (6), the improved chi-square statistic formula becomes:

$$\langle\langle MATH_{38}\rangle\rangle$$

2.2 ECE Analysis and Weighting

The ECE method considers both the correlation between features and categories and the difference between feature frequency and category frequency. However, it has clear shortcomings. From equation (5), if $\langle\langle MATH_{39}\rangle\rangle$ is large and $\langle\langle MATH_{40}\rangle\rangle$ is small, feature t has a significant impact on classification, resulting in a large ECE(t) value. This demonstrates that the ECE method does not consider the inter-class distribution of features in the dataset, causing the algorithm to favor high-frequency features with weak discriminative ability [10].

To address these deficiencies, this paper comprehensively considers both feature presence/absence and inter-class occurrence proportions, using term frequency $\langle\langle MATH_{41}\rangle\rangle$ as a weighting scheme to evaluate the information content of a feature. The normalized weight calculation formula is:

$$\langle\langle MATH_{42}\rangle\rangle$$

Combining equations (5) and (8), the weighted expected cross-entropy formula becomes:

$$\langle\langle MATH_{43}\rangle\rangle$$

where $\langle\langle MATH_{44}\rangle\rangle$ reflects the distance between the probability distribution of text category i and the conditional probability distribution of documents belonging to class i containing feature t. A larger expected cross-entropy value indicates greater impact on text classification.

2.3 pCHIωECE Feature Selection Method

Through analyzing the characteristics and shortcomings of CHI and ECE, this paper optimizes both feature selection methods individually. By integrating these two improved methods, we propose a feature selection method based on combining improved CHI and weighted ECE (pCHIωECE). The design flow of pCHIωECE is shown in [Figure 1: see original paper].

The improved CHI (pCHI) method reduces interference from low-frequency features in text classification and removes negative correlation influences, giving pCHI better dimensionality reduction capability compared to CHI. The ωECE method mitigates ECE' s reliance on high-frequency features with weak discriminative ability, thereby improving classification effectiveness to some extent. The pCHIωECE method combines the characteristics of both CHI and ECE, not only alleviating the low-frequency word deficiency but also selecting features that appear frequently in specific categories, resulting in better classification performance and stability. The pCHIωECE calculation formula is:

$$\langle\langle MATH_{45} \rangle\rangle$$

---

3 Experiments

3.1 Dataset and Experimental Setup

The experimental environment consists of a PC with Windows 10 x64 operating system, Intel(R) CoreTM i5-5250U @ 1.6 GHz processor, and 4 GB memory, using Python 3.6 as the development tool. We implemented five feature selection methods (the referenced, improved, and proposed methods) by calling Python' s Sklearn module, and selected the Naive Bayes (NB) classifier for classification.

The dataset source is the Fudan University Chinese corpus, containing 20 categories with 9,833 documents. The number of documents per category is as follows: Space (642), Energy (33), Electronics (28), Communication (27), Computer (1,358), Mining (34), Transportation (59), Art (742), Environment (1,218), Agriculture (1,022), Economy (1,601), Law (52), Medical (53), Military (76), Politics (1,026), Sports (1,254), Literature (34), Education (61), Philosophy (45), and History (468). The distribution across categories is highly imbalanced. In our experiments, 80% of the data served as the training set and 20% as the test set.

3.2 Evaluation Metrics

We employ standard performance evaluation metrics for text classification [11]: precision, recall, and F1-score. For text classification problems, the combination of actual and predicted categories yields four cases: true positive (TP), false

positive (FP), true negative (TN), and false negative (FN). The confusion matrix for classification results is shown in .

Precision is defined as:

$$\langle\langle MATH_{46}\rangle\rangle$$

Recall is defined as:

$$\langle\langle MATH_{47}\rangle\rangle$$

F1-score is the harmonic mean of precision and recall, defined as:

$$\langle\langle MATH_{48}\rangle\rangle$$

To comprehensively consider both precision and recall, this paper uses the F1-score as the primary evaluation metric.

The precision rates of traditional CHI and ECE feature selection methods are compared in [Figure 2: see original paper]. The results show that both methods exhibit significant fluctuations in precision across different feature quantities, with ECE generally achieving higher precision. When the number of features exceeds 650, CHI' s precision becomes more stable.

[Figure 3: see original paper] compares the precision of the improved CHI method (pCHI) with the original CHI method. The pCHI method introduces an adjustment factor and removes negative correlation cases between features and categories. The results show that pCHI' s precision trend correlates with CHI to some extent, but pCHI consistently achieves higher precision than traditional CHI, validating its correctness and advantages.

[Figure 4: see original paper] compares the precision of the weighted ECE method (ωECE) with the original ECE method. ωECE considers both feature presence and inter-class occurrence proportion for weighting. The results show that ωECE exhibits more stable precision trends than ECE, with higher precision values, confirming its correctness and advantages.

[Figure 5: see original paper] compares the precision of pCHIωECE with CHI, pCHI, ECE, and ωECE. The pCHIωECE method achieves higher precision than the other four methods. While traditional CHI and ECE methods have relatively low accuracy, their performance improves as the number of features increases. Both pCHI and ωECE show slight improvements over their original versions, with comparable dimensionality reduction capabilities. The results indicate that imbalanced inter-class distribution in the text dataset affects classification outcomes. Higher accuracy when the number of features is below 700 demonstrates that pCHIωECE has better dimensionality reduction capability.

The method maintains high overall stability, improves classification efficiency under imbalanced dataset conditions, and achieves superior overall precision.

The F1-scores of the five feature selection methods using Naive Bayes classification are shown in [Figure 6: see original paper]. Comparative analysis reveals that CHI and ECE methods show no clear correlation with feature quantity. Both pCHI and ωECE achieve higher F1-scores than their original versions, while pCHIωECE attains the highest F1-score among all methods.

Based on both precision and F1-score results, the proposed feature selection method based on combining improved CHI and weighted ECE demonstrates superior classification performance compared to CHI and ECE methods.

---

4 Conclusion

This paper first analyzed the characteristics and shortcomings of CHI and ECE feature selection methods. We improved CHI by introducing adjustment factors and removing negative correlation influences, and optimized ECE using a weighting approach. Building upon these improvements, we proposed a feature selection method combining improved CHI and weighted ECE. Comparative experiments on an imbalanced dataset analyzed the precision and F1-scores of different methods, qualitatively demonstrating that pCHI, ωECE, and pCHIωECE all achieve varying degrees of improvement in classification effectiveness. These results validate that the pCHIωECE method can enhance text classification accuracy while providing better stability.

References

[1] Aggarwal C C, Zhai ChengXiang. A survey of text classification algorithms [M]// Mining Text Data. Boston: Springer, 2012: 163-222.

[2] Cai Jie, Luo Jiawei, Liang Cheng, et al. A novel information theory-based ensemble feature selection framework for high-dimensional microarray data [J]. International Journal of Performability Engineering, 2017, 13 (5): 742-752.

[3] Vijayan V K, Bindu K R, Parameswaran L. A comprehensive study of text classification algorithms [C]// Proc of International Conference on Advances in Computing, Communications and Informatics. 2017: 1109-1113.

[4] Mao Yong, Zhou Xiaobo, Xia Zheng, et al. A survey for study of feature selection algorithms [J]. Pattern Recognition and Artificial Intelligence, 2007, 20 (2): 211-218.

[5] Zhang Qun, Wang Hongjun, Wang Lunwen. Algorithm of text feature selection based on vocabulary attribute clustering [J]. Application Research of Computers, 2017, 34 (2): 369-372, 377.

[6] Rehman A, Javed K, Babri H A. Feature selection based on a normalized difference measure for text classification [J]. Information Processing & Management, 2017, 53 (2): 473-489.

[7] Gao Baolin, Zhou Zhiguo, Yang Wenwei, et al. Feature selection method based on combination of category and improved CHI [J]. Application Research of Computers, 2018, 35 (6): 1660-1662.

[8] Wang Haijuan, Han Lixin, Zhen Zhilong. Feature selection based on term weight for text categorization [J]. Computer Engineering and Design, 2010, 31 (5): 1149-1151.

[9] Shan Lili, Liu Bingquan, Sun Chengjie. Comparison and improvement of feature selection method for text categorization [J]. Journal of Harbin Institute of Technology, 2011, 43 (S1): 319-324.

[10] Du Tongsen. Research on feature selection and feature weighting algorithm in text categorization [D]. Beijing: Beijing University of Posts and Telecommunications, 2014.

[11] Wang Zhen, Qiu Xiaohui. An improved text feature selection method mixed CHI and MI [J]. Computer Technology and Development, 2018, 28 (4): 87-90, 94.

[12] Liu Jinghua, Lin Yaojin, Lin Menglei, et al. Feature selection based on quality of information [J]. Neurocomputing, 2017, 225: 11-22.

[13] Meng Jiana, Lin Hongfei, Li Yanpeng. Application of feature selection method to text categorization based on feature contribution degree [J]. Journal of Dalian University of Technology, 2011, 51 (4): 611-615.

[14] Qiu Yunfei, Wang Wei, Liu Dayou, et al. CHI feature selection method based on variance [J]. Application Research of Computers, 2012, 29 (4): 1304-1306.

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