Postprint: Application of Shuffled Frog Leaping Algorithm in Feature Selection Optimization for Text Classification

Authors: Lu Yonghe, Jinghuang Chen

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: As text data contains numerous redundant terms unrelated to classification, the shuffled frog leaping algorithm is introduced to optimize feature selection and improve classification accuracy.

Methods: CHI and IG are respectively employed to pre-select feature sets of varying dimensions, followed by the introduction of an improved shuffled frog leaping algorithm for secondary optimization of the pre-selected feature sets. Each frog' s position represents a feature selection rule, with classification accuracy serving as the algorithm' s fitness function. SVM and KNN classifiers are utilized for calculating classification accuracy in the experiments.

Results: The introduced improved frog leaping algorithm achieves superior classification performance compared to CHI and IG, with a maximum improvement of 12%.

Limitations: Overfitting occurs under a few feature dimensions.

Conclusion: The feature selection optimization method combining feature term pre-selection and improved frog leaping algorithm can effectively eliminate interference from noisy feature terms, thereby enhancing text classification accuracy.

Full Text

Optimizing Feature Selection for Text Classification with the Shuffled Frog Leaping Algorithm

Lu Yonghe, Chen Jinghuang
(School of Information Management, Sun Yat-sen University, Guangzhou 510006, China)

Abstract:
[**Objective**] Due to the presence of numerous irrelevant redundant terms in textual data, this paper introduces the Shuffled Frog Leaping Algorithm (SFLA) to optimize feature selection and improve classification accuracy. [**Methods**] We first used CHI and IG methods to pre-select feature sets of varying dimensions, then applied an improved SFLA to perform secondary optimization on these preselected feature collections. Each frog' s position represents a feature selection rule, with classification accuracy serving as the algorithm's fitness function. Both SVM and KNN classifiers were employed to calculate classification accuracy in the experiments. [**Results**] The improved SFLA achieved better classification performance than CHI and IG alone, with a maximum improvement of 12%. [**Limitations**] Overfitting occurred in a small number of feature dimensions. [**Conclusions**] The proposed method combining feature term preselection with the improved SFLA can effectively eliminate interference from noisy feature terms, thereby enhancing text classification accuracy.

Keywords: Feature Selection; Text Classification; Shuffled Frog Leaping Algorithm

---

Text classification serves as a fundamental technology for information mining, natural language processing, and information retrieval, attracting considerable scholarly attention. The technology has evolved from manual classification to machine learning-based automatic classification, significantly improving both quality and efficiency. However, textual data often exhibits high dimensionality, sparsity, and multi-label characteristics that affect classification performance, making text feature selection optimization a research focus. In the Vector Space Model (VSM), not all features in the original feature set are necessary for classification learning; some noisy features not only increase dimensionality but also degrade overall classification effectiveness. Therefore, dimensionality reduction of the feature set is essential.

This paper employs the Shuffled Frog Leaping Algorithm (SFLA), which has seen limited application in text processing, and improves its encoding rules and individual evolution mechanisms for text feature selection optimization. Experimental results demonstrate the effectiveness of this approach.

2.1 Traditional Text Feature Selection Methods

The text classification process primarily includes text preprocessing and segmentation, text representation, feature selection, weight calculation, and classification. Text representation mainly adopts the VSM model, where preprocessed texts yield extremely high-dimensional feature sets with sparse distributions, representing each document as a high-dimensional vector. This imposes significant computational burdens on classifiers, making feature selection crucial for reducing vector space dimensionality while improving classification efficiency and accuracy. Current feature selection methods include Document Frequency

(DF), Chi-square test (CHI), Information Gain (IG), and Mutual Information (MI). Experimental evidence indicates that CHI offers good classification performance but high computational overhead. In English text classification, CHI and IG perform best, with DF being comparable, while MI performs relatively poorly. In Chinese text classification, CHI performs best, followed by IG, with MI performing poorly and DF achieving moderate results.

However, traditional methods like CHI and IG select features through mathematical models that identify terms with good discriminative power and representativeness, without considering inter-feature relationships or the overall impact of redundant terms from a textual perspective. Therefore, this paper introduces an improved SFLA to perform secondary optimization on preselected feature sets, leveraging the algorithm' s strong search capabilities to obtain relatively low-dimensional yet high-precision feature collections that ultimately improve classification results.

2.2 Feature Selection Optimization with Swarm Intelligence Algorithms

Recent years have witnessed growing applications of swarm intelligence algorithms in text feature selection with notable effectiveness. The general approaches fall into two categories:

First, using swarm intelligence algorithms directly for feature selection without traditional methods. Representative work includes: Tabakhi et al.' s UFSACO method, which introduces Ant Colony Optimization (ACO) into unsupervised feature selection, considering feature correlations to remove redundant terms and achieve dimensionality reduction with improved classification performance; Liu Yanan' s application of Genetic Algorithm (GA)-based feature selection in KNN classification with dynamic K-value acquisition; and Liu Kui' s text feature selection model based on Invasive Weed Optimization, which provides opportunities for low-weight terms while preserving advantages of high-weight terms to comprehensively improve selection coverage and accuracy.

Second, combining swarm intelligence algorithms with traditional feature selection methods by first obtaining preselected feature sets through conventional approaches, then applying swarm intelligence algorithms for refinement to obtain high-precision collections. Representative studies include: Uguz's two-stage method using IG followed by GA or Principal Component Analysis (PCA) to remove irrelevant terms; Javed et al.' s combination of BNS and IG preselection with Markov Blanket Filter (MBF) for secondary screening; and Lu et al.' s use of CHI preselection with six improved Particle Swarm Optimization (PSO) variants, demonstrating that asynchronous improved PSO achieved the best classification performance.

This paper adopts the second approach, combining SFLA with traditional methods through feature term preselection followed by improved binary SFLA refinement to obtain high-precision feature sets and ultimately improve classification

effectiveness.

2.3 Shuffled Frog Leaping Algorithm

The Shuffled Frog Leaping Algorithm, proposed by Eusuff et al., is a collaborative search swarm intelligence algorithm combining characteristics of Memetic Algorithms (MA) and Particle Swarm Optimization. It inherits both the genetic properties of MA and the social information sharing features of PSO, offering simple and reasonable flow, few parameters, fast convergence, and strong global optimization capabilities.

Inspired by frog foraging behavior, SFLA involves a population P of N frogs searching for optimal food sources in an S-dimensional constrained space. Each frog i' s position is represented as $X_i = (x_{i1}, x_{i2}, ..., x_{iS})$, where S denotes spatial dimensionality and X_i represents a feasible solution vector for optimization problems. The fitness $F(X_i)$ of each frog' s current position is calculated, followed by descending sorting and recording of the global best position X_g. The population is divided into n memplexes, each containing m frogs, using a round-robin grouping rule. Each memplex records its local best solution X_b and worst solution X_w. Intra-memplex evolution proceeds as follows:

$$D = \text{rand}() \times (X_b - X_w) \quad (1)$$

where D represents the step size for each jump, rand() is a random number between 0 and 1, and X'_w denotes the position after jumping:

$$X'_w = X_w + D \quad (2)$$

If X'_w' s fitness $F(X'_w)$ is better than X_w' s fitness $F(X_w)$, X'_w replaces X_w for the next evolution; otherwise, X_g replaces X_b in equation (1) to recalculate X'_w. If X'_w improves upon X_w, it replaces X_w; otherwise, a random X'_w is generated to replace X_w. When all memplexes reach the maximum evolution count L, all frogs are reshuffled, re-sorted by fitness, and the global best solution X_g is updated to form the next generation until reaching maximum iterations T or satisfying termination conditions.

SFLA has been applied to water resource network optimization, bridge deck repair, dynamic optimal power flow calculation in wind power systems, distributed wind generator planning, and speech recognition. However, literature shows limited application in text information processing. Xu Fang improved traditional SFLA and combined it with K-means and FCM for web text clustering, improving precision. Wei Jianxing et al. similarly combined SFLA with K-means to enhance clustering performance. Sun et al. used SFLA directly as a classification algorithm with LDA for feature selection, improving web text classification accuracy. Overall, SFLA remains underutilized in text processing. This paper

attempts to improve SFLA and combine it with traditional feature selection methods, validating its effectiveness and feasibility through experiments.

3. Text Feature Selection Optimization Based on SFLA

3.1 Algorithm Improvements (1) Encoding Rules

Since text feature selection optimization is essentially a combinatorial optimization problem, SFLA requires binary encoding modification. Each frog's position represents a feature selection rule, where each dimension corresponds to a feature term with two possible outcomes: selected (1) or not selected (0). Thus, each solution vector (frog position) can be represented as $X_i = \{x_{i1}, x_{i2}, ..., x_{iS}\} \in \{0,1\}^S$, where x_{ij} denotes the j-th component of the i-th solution vector, taking values of 0 or 1. If $x_{ij} = 1$, the j-th feature term is selected; if $x_{ij} = 0$, it is not selected.

(2) Individual Evolution Modification

Due to binary encoding, standard SFLA' s evolution mechanism (equations (1) and (2)) becomes inapplicable. The following improvements enable better suitability for text feature selection optimization, as illustrated in [Figure 1: see original paper].

First, we identify the intersection G of selected feature terms between the memplex' s best solution X_b and worst solution X_w (i.e., all components where both X_b and X_w equal 1). Treating X_b and X_w as sets:

$$G = X_b \cap X_w$$

Next, we calculate each frog' s jump step D_{new}:

$$D_{new} = R_1 \cup R_2$$

where $(X_b - X_w)$ and $(X_w - X_b)$ represent set difference operations. r_1 and r_2 are random integers between 0 and 100. R_1 contains the top $r_1\%$ of feature elements from $(X_b - X_w)$, while R_2 contains the top $r_2\%$ from $(X_w - X_b)$. The union D_{new} constitutes the step size for a frog' s jump. For example, when $r_1 = 20$, $r_2 = 40$, $(X_b - X_w)$ contains 100 elements, and $(X_w - X_b)$ contains 200 elements, R_1 selects the top $100 \times 20\% = 20$ features and R_2 selects the top $200 \times 40\% = 80$ features, forming D_{new} with $20 + 80 = 100$ features. The updated position after jumping becomes:

$$X'_w = G \cup D_{new}$$

This modification is based on the following rationale: The intersection G preserves "common features" between X_b and X_w, allowing new individuals to

"inherit" these shared features while continuing evolution toward better positions. The step size D_{new} randomly "inherits" proportions of "unique" features from both X_b and X_w, enabling directional evolution. Since candidate feature sets are pre-screened by CHI or IG and sorted by descending scores, selecting top-ranked features ensures representativeness.

(3) Maximum Step Size D_{max} Modification

With the modified step size calculation, the maximum step size D_{max_new} must also be redefined. We introduce a "difference degree"(diff) variable measuring the proportion of differing components between the new individual X'_w and original X_w. D_{max_new} represents the maximum allowed difference degree. For instance, if $X'_w = \{0, 1, 1, 1, 0, 0\}$ and $X_w = \{1, 0, 1, 1, 0, 1\}$ differ at dimensions 1, 2, and 6, then diff $= (3/6) \times 100\% = 50\%$. The difference degree variable quantifies the disparity proportion between binary-encoded frog individuals, analogous to step size in standard SFLA, but adapted for the modified binary step calculation.

3.2 Parameter Settings The improved binary SFLA requires five parameters: frog population size N, number of memplexes n, maximum step size D_{max}, intra-memplex evolution count L, and total iterations T. Parameter settings significantly impact performance.

Population size N refers to the total number of frogs (initial solution vectors). While generally correlated with problem complexity, computational overhead considerations led to setting N = 20. The number of memplexes n depends on frogs per memplex (m); we set n = 5, yielding m = 4 frogs per memplex. The maximum step size D_{max} controls global search capability by limiting the difference degree between new and original individuals; we set $D_{max} = 45\%$, meaning component differences cannot exceed 45%. Intra-memplex evolution count L determines evolution rounds per memplex, set to L = 10 due to computational costs. Total iterations T correlates with problem complexity; we set T = 10.

3.3 Fitness Function Swarm intelligence algorithms calculate individual fitness based on optimization objectives. This paper aims to reduce feature set dimensionality while improving classification accuracy. Therefore, classification accuracy measures each frog' s position quality, guiding frogs to "leap" toward higher accuracy:

$$\text{Fitness} = \frac{\text{Number of correctly classified test texts}}{\text{Total number of texts in test set}}$$

3.4 Algorithm Design The improved SFLA-based text feature selection optimization algorithm proceeds as follows:

Input: Training text set TR, test text set A, desired preselected feature count S (feature space dimension), initial frog count N, memplex count n, maximum

step size D_{max}, intra-memplex evolution limit L, total iterations T.

Output: Feature set after SFLA secondary optimization.

1. Preprocess training set TR, then apply CHI and IG for feature preselection to obtain candidate feature sets.
2. Randomly initialize each frog' s position dimensions with {0,1} values, where 1 indicates feature selection and 0 indicates exclusion.
3. Calculate each frog' s fitness (classification accuracy). Features with value 1 form the representation for test set A, constructing feature vectors for accuracy calculation.
4. Execute the improved SFLA until iteration count reaches T or termination conditions are met, then output the global best solution X_g and corresponding selected features.

The algorithm flow is illustrated in [Figure 2: see original paper].

4. Experiments

The experiments comprise two parts: (1) Direct application of traditional CHI or IG selected features for classification; (2) Introduction of SFLA for secondary optimization to obtain high-precision feature sets for classification, as shown in [Figure 3: see original paper].

In the direct application process, datasets TR and B are used to calculate classification accuracy for original feature sets. The SFLA optimization process requires both training and test sets for fitness calculation, thus using TR and A as the modeling dataset. After obtaining high-precision feature sets, test set B evaluates performance to verify generalization. Using separate test sets prevents overfitting to set A and reduces computational overhead during optimization (set A is smaller, comprising 15% randomly sampled from each category of set B).

Experiments use both English and Chinese datasets: Reuters-21578 (Experiment 1) and a corpus from the Intelligent Information Processing Laboratory of Sun Yat-sen University' s School of Information Management (Experiment 2).

The experimental environment was 32-bit Windows 10, 4GB RAM, i5-2400 processor, with Java implementation. Text preprocessing used the Lucene package, and Chinese word segmentation employed ICTCLAS from the Institute of Computing Technology, Chinese Academy of Sciences. CHI and IG performed feature preselection, TF-IDF calculated weights, and SVM and KNN served as classifiers.

4.1 Reuters-21578 Corpus Experiment Reuters-21578 contains 8 categories: acq, crude, earn, grain, interest, money-fx, ship, and trade. The large test set and training set were split at a 1:2.5 ratio, with specific distributions shown in .

Procedure: 1. Using TR and B, CHI preselected 12 feature sets (100-1200

dimensions at 100-interval steps: CHI_{100}-CHI_{1200}), calculating accuracies P_{CHI}. 2. Improved binary SFLA refined these 12 sets using TR and A as the training dataset, outputting optimized high-precision sets. 3. Using TR and B, accuracies P_{CHI_SFLA} were calculated for optimized sets. 4. Steps 1-3 repeated using IG, yielding P_{IG} and P_{IG_SFLA}. 5. Accuracies were compared across 12 dimensions. 6. All accuracies were grouped into P_{old} (without SFLA) and P_{new} (with SFLA) for paired sample t-testing.

SVM Classifier Results: shows classification accuracies. Figures [Figure 4: see original paper] and [Figure 5: see original paper] plot CHI and IG results respectively. With SVM on Reuters-21578, the improved SFLA secondary optimization consistently outperformed traditional CHI and IG, with improvement margins increasing with dimensionality.

KNN Classifier Results: shows accuracies, plotted in [Figure 6: see original paper] and [Figure 7: see original paper]. With KNN, SFLA secondary optimization generally outperformed CHI and IG. At 400 dimensions, IG_{SFLA} matched IG accuracy but with reduced dimensionality, indicating IG' s 400-dimensional set contained classifiable irrelevant terms.

4.2 Laboratory Corpus Experiment The laboratory corpus (collected by Sun Yat-sen University' s Intelligent Information Processing Laboratory) contains 13 categories. Eight categories with sufficient texts were selected: education, entertainment, event, finance, game, occultism, sport, and technology. Each category contributed 200 texts to training set TR (1,600 total) and 200 texts to test set B (1,600 total). An additional test set A contained 160 texts (20 per category). After preprocessing, TR yielded 52,794 unique features.

SVM Classifier Results: shows accuracies, plotted in [Figure 8: see original paper] and [Figure 9: see original paper]. The improved SFLA outperformed CHI and IG, with both achieving peak accuracy at 1,000 dimensions. Improvement margins reached approximately 7% for CHI_{SFLA} at 400 dimensions and 9% for IG_{SFLA} at 300 dimensions. Optimization was particularly effective when traditional methods produced lower baseline accuracies.

KNN Classifier Results: and Figures [Figure 10: see original paper] and [Figure 11: see original paper] show that CHI_{SFLA} outperformed CHI (though marginally at 100 and 1,000 dimensions), while IG_{SFLA} significantly outperformed IG, achieving 12% improvement at 1,000 and 1,100 dimensions.

4.3 Paired Sample T-Test All accuracy data were grouped into P_{old} (without SFLA) and P_{new} (with SFLA) for paired sample t-testing in SPSS. shows Sig. $= .000 < 0.01$, indicating significant differences at the 99% confidence level. SFLA-based feature optimization significantly improved text classification accuracy.

5. Conclusion

Experiments on both corpora demonstrate that the improved SFLA-based text feature selection optimization algorithm outperforms traditional CHI and IG methods. The improvement stems from SFLA' s iterative optimization and convergence properties, which remove noisy feature terms that traditional statistical methods overlook, thereby enhancing classification accuracy.

This paper introduces SFLA—rarely applied in text processing—to feature selection optimization. Comparative experiments show the proposed method achieves higher accuracy by reducing noise term interference. However, parameter settings were determined through small-scale tests. Future work will optimize SFLA parameters to approach optimal solutions more closely and achieve even better feature sets and classification performance.

References

[1] Pang Guansong, Jiang Shengyi. Text Automatic Classification Technology Research [J]. Information Studies: Theory & Application, 2012, 35(2): 123-128.

[2] Wu Ke. A Study on Text Categorization Based on Machine Learning [D]. Shanghai: Shanghai Jiaotong University, 2008.

[3] Wu Jianjun, Kang Yaohong. Comparison and Improvement of Feature Selection for Text Categorization [J]. Journal of Zhengzhou University: Natural Science Edition, 2007, 39(2): 110-113.

[4] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1997: 412-420.

[5] Fu Fa. Comparison of Feature Selection in Chinese Text Categorization [J]. Modern Computer, 2008(6): 43-45.

[6] Tabakhi S, Moradi P, Akhlaghian F. An Unsupervised Feature Selection Algorithm Based on Ant Colony Optimization [J]. Engineering Applications of Artificial Intelligence, 2014, 32: 112-123.

[7] Liu Yanan. Research of Feature Extraction Technology in KNN Text Classification Based on the Genetic Algorithm [D]. Beijing: China University of Petroleum, 2011.

[8] Liu Kui. An Invasive Weed Optimization Algorithm for Text Feature Selection [D]. Chongqing: Southwest University, 2013.

[9] Uguz H. A Two-stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm [J]. Knowledge-Based Systems, 2011, 24(7): 1024-1032.

[10] Javed K, Maruf S, Babri H A. A Two-stage Markov Blanket Based Feature Selection Algorithm for Text Classification [J]. Neurocomputing, 2015, 157: 91-104.

[11] Lu Y, Liang M, Ye Z, et al. Improved Particle Swarm Optimization Algorithm and Its Application in Text Feature Selection [J]. Applied Soft Computing, 2015, 35(C): 629-636.

[12] Eusuff M M, Lansey K E. Optimization of Water Distribution Network Design Using the Shuffled Frog Leaping Algorithm [J]. Journal of Water Resources Planning and Management, 2003, 129(3): 210-225.

[13] Cui Wenhua, Liu Xiaobing, Wang Wei, et al. Survey on Shuffled Frog Leaping Algorithm[J]. Control and Decision, 2012, 27(4): 481-486, 493.

[14] Elbehairy H, Elbeltagi E, Hegazy T, et al. Comparison of Two Evolutionary Algorithms for Optimization of Bridge Deck Repairs [J]. Computer-Aided Civil and Infrastructure Engineering, 2006, 21(8): 561-572.

[15] Chen Gonggui, Li Zhihuan, Chen Jinfu, et al. SFLA Algorithm Based Dynamic Optimal Power Flow in Wind Power Integrated System [J]. Automation of Electric Power Systems, 2009, 33(4): 25-30.

[16] Zhang Shenxi, Chen Kai, Long Yu, et al. Distributed Wind Generator Planning Based Shuffled Frog Leaping Algorithm [J]. Automation of Electric Power Systems, 2013, 37(13): 76-82.

[17] Yu Hua, Huang Chengwei, Jin Yun, et al. Speech Emotion Recognition Based on Modified Shuffled Frog Leaping Algorithm Neural Network [J]. Signal Processing, 2010, 26(9): 1294-1299.

[18] Xu Fang. Research on Web Text Cluster Algorithm Based on Shuffled Frog-leaping Algorithm [D]. Wuxi: Jiangnan University, 2013.

[19] Yu Jianxing, Cui Donghua, Ning Xiaoqing. Application of Shuffled Frog-leaping Algorithm to Web's Text Cluster Technology [J]. Computer Development & Applications, 2011, 24(5): 35-37.

[20] Sun X, Wang Z. An Efficient Document Categorization Algorithm Based on LDA and SFL [C]//Proceedings of the 2008 International Seminar on Business and Information Management. IEEE, 2008: 113-115.

[21] NLPIR Chinese Word Segmentation System [EB/OL]. [2016-03-17]. http://ictclas.nlpir.org.

[22] Lu Yonghe, Peng Yanhong. The Classification System Construction for Internet Information both Practical and Scientific[J]. Library and Information, 2015(3): 118-124.

Author Contributions

Lu Yonghe: Conceived research ideas and experimental suggestions, revised the manuscript.
Chen Jinghuang: Analyzed data, designed and implemented algorithms, conducted experiments, wrote and revised the final manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data available in the journal's online version at http://www.infotech.ac.cn:
[1] Lu Yonghe, Chen Jinghuang. Experimental datasets.rar. Selected text collections from Reuters-21578 and Sun Yat-sen University Intelligent Laboratory corpora.
[2] Lu Yonghe, Chen Jinghuang. Preselected feature sets.rar. Feature term sets preselected via CHI and IG.
[3] Lu Yonghe, Chen Jinghuang. Optimized feature sets.rar. Refined feature term sets selected by improved SFLA.
[4] Lu Yonghe, Chen Jinghuang. Classification results.xlsx. Text classification accuracies obtained using SFLA-refined feature sets.

Received: September 30, 2016
Revised: December 12, 2016

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