Dimensionality Reduction by Feature Co-Occurrence based Rough Set

Lei La^{a,*}, Qimin Cao^b, and Ning Xu^b

^a*School of Information Technology & Management, University of International Business and Economics, Beijing, 100029, China*

^b*Library, China University of Political Science and Law, Beijing, 100088, China*

Abstract

Feature selection is the key issue of unstructured data mining related fields. This paper presents a dimensionality reduction method which uses a rough set as the feature selection tool. Different from previous rough set based classification algorithm, it takes feature co-occurrence into account when make attribution reduction to get a more accurate feature subset. The novel method called Feature Co-occurrence Quick Reduction algorithm is in this article. Experimental results show it has a high efficiency in dimensionality reduction—time consumption by approximately 23% less than traditional rough set based dimensionality reduction methods. Moreover, classification based on the feature set selected by Feature Co-occurrence Quick Reduction algorithm is more precise. The proposed algorithm is helpful to us for refining knowledge from massive unstructured data.

Keywords: dimensionality reduction; feature co-occurrence; rough set; time consumption; unstructured data

(Submitted on October 23, 2018; Revised on November 20, 2018; Accepted on December 28, 2018)

© 2019 Totem Publisher, Inc. All rights reserved.

1. Introduction

Feature selection and feature extraction are collectively referred to as dimensionality reduction. It is the foundation of many data mining tasks. Especially in the textual data set, the dimensionality of features is often tens of thousands. This could easily lead to the curse of dimensionality and feature sparsity problems. Feature dimension of a middle-size document set may reach 10^4 or 10^5 [1] and extremely increases the computational and runtime complexity of the task. This is the so called curse of dimensionality. Feature sparsity means the occurrence probability of a feature in a certain document which belonged to the document set is very low. In other words, in the vector space, most components of a text are zero-vectors. Feature sparsity would greatly reduce the accuracy of classification and waste the processing time [2].

To solve the problems mentioned above, some scholars proposed selection based algorithms such as Principal Component Analysis (PCA) [3-4] and Locally Linear Embedding (LLE) [5-6]. In text data mining, these methods use words as the features to save computational cost. On the other hand, some researchers commit to extracting concepts or topics as the features. Most classic feature extraction methods are Probabilistic Semantic Indexing (PSI) [7] and Latent Dirichlet Allocation (LDA) [5]. Generally speaking, feature selection has higher efficiency and is much more popular in dimensionality reduction. The schematic of feature selection is shown in Figure 1.

Rough Set has proven to be a powerful tool for data classification and especially for text classification [8]. Inspired by their (Duoqian Miao, Qiguo Duan, Hongyun Zhang, et al.) work, we proposed a Rough Set based dimensionality reduction for text categorization in this article.

In this paper, we take co-occurrence features into account when we make a decision table reduction and propose a Feature Co-occurrence Quick Reduction (FCQR) algorithm. FCQR is introduced into a rough set for the document feature selection. Experimental results show the novel method has a higher dimensionality reduction efficiency. Moreover, the high

* Corresponding author.

E-mail address: lalei1984@aliyun.com

classification accuracy based on features selected by the novel method reveals it has a more powerful ability to select the most representative features than many other algorithms.

Figure 1. Schematic of feature selection

The method presented in this paper is a powerful tool in machine learning and data mining related fields since it gives an ideal solution of dimensionality reduction. As we know, dimensionality reduction is a basis problem in data mining associated topics such as knowledge discovery in database, natural language processing, image recognition and so on. Therefore, this paper is devoted to making machine learning more efficient through rough set theory.

2. Review of Routh Set

Rough set theory was proposed by Polish mathematician Pawlak [9]. It is an analysis tool for an imprecise, incomplete and inconsistent knowledge set.

Information system U can be expressed as

$$U = (W, C, X, h) \quad (1)$$

In Equation (1) W is the collection of objects, it also can be called universe of discourse, C is the set of attributions, $X = \bigcup_{\alpha \in C} X_\alpha$, is the range of attribution α , $h: W \times C \rightarrow X$ is an information function who assigned the property value of each object z in W . In other words, for $z \in W$ and $c \in C$, $h(z, c) \in X_c$. Given information system $U = (W, C, X, h)$, let $Z \subseteq W$ as a group of objects and $D \subseteq C$ as a group of attributions. As a result, the lower approximation relative to D of Z is:

$$\underline{D}(Z) = \{z \mid z \in W, [z]_D \subseteq Z\} \quad (2)$$

To solve Equation (2), the upper approximation of Z relative to D is:

$$\bar{D}(Z) = \{z \mid z \in W, [z]_D \cap Z \neq \emptyset\} \quad (3)$$

In Equation (3), in an information system, Object set $Z \subseteq W$ called concept. It is an accurate set when the lower and upper approximations are equal, otherwise, it is a rough set. Where lower approximation is called the positive region in the concept, the area outside the upper approximation is called the negative region and the difference between the lower and upper approximation is called the *boundary*. S_E is the empty set.

3. Decision Table and its Reduction

If the attribution set C can be divided into condition attribution set E and decision attribution F , where $E \cup F = C$ and $E \cap F = S_E$, the information system can be called a decision system or decision table. Usually, F contains only one attribution $\{f\}$. The positive region of decision attribution f relative to D is:

$$POS_D(f) = \bigcup_{Z \in W / IND(\{f\})} \underline{DZ} \quad (4)$$

In Equation (4), $IND(f)$ is an undistinguishable equivalence relation:

$$IND(f) = \{(x, y) : x, y \in W \\ \forall f \in F, h(x, f) = h(y, f)\} \quad (5)$$

Applying rough set theory as Equation (5) into knowledge acquisition commonly needs learning from examples at first. Therefore, the definition of a decision table is very important. Similar with a knowledge table, a decision table contains a large number of instance records. In addition, it contains decision attributions which don't exist in a knowledge table. The positive region of decision attribution f represents the approximation quality of $D \subseteq C$ relative to f . For decision-making of the original data, not all the attributions in C are necessary. The purpose of knowledge acquisition is analysing the case database to discover which attributions are important for decision and which attributions are less important or redundant. Reduction is finding the minimal condition attribution subset when maintain the approximate amount stable. After the introduction of decision table, assuming the decision attribution set is F , the calculation method of dividing positive region is:

$$POS_C(F) = \bigcup_{Z \in W / F} \underline{CZ} \quad (6)$$

In Equation (6), the definition of degree of dependence between attribution set C and decision subset F is:

$$\delta_C(F) = \frac{|POS_C(F)|}{|W|} \quad (7)$$

In Equation (7), for $\forall x \in F$, the significance of attribution x for decision subset F is:

$$\delta_x(F) = \delta_C(F) - \delta_{C-\{x\}}(F) \quad (8)$$

Hitherto, we can make a decision table reduction according to the importance of attribution for decision making. As shown above in Equation (8), it is a concise way.

4. Review of Routh Set

4.1. Text Feature Space Modelling by Rough Set

According to the training document collection $D = \{d_1, d_2, \dots, d_n\}$, we can get the candidate feature set $T = \{t_1, t_2, \dots, t_m\}$. The value of each document in features is 0 or 1 which means the features appear or do not appear in the text. Use document set D as the universe of discourse, candidate feature set T as the condition attribution subset E , and text category set C as the decision attribution subset F , where $C = \{c_1, c_2, \dots, c_l\}$. The decision table for text categorization can be constructed as Table 1.

Table 1. Decision table for text categorization

D	T				C
	t_1	t_2	$\dots$	t_m	
d_1	1	1	$\dots$	0	c_1
d_2	0	1	$\dots$	0	c_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
d_{n-1}	0	1	$\dots$	0	c_{l-1}
d_n	1	0	$\dots$	1	c_l

Table 1 has three characteristics which make it difficult to be classified, and it may need some further mining, the difficulties are as follows.

(1) Large scale of condition attribution set, in another word, the dimension of document set is extremely high.

(2) Sparsity. That because the number of features in each text is relatively small.

(3) Significantly uneven distribution of non-zero attributions. It means a particular feature have different weight in different document

4.2. Decision Table Reduction

In rough set based text mining tasks, feature selection is equivalent to decision table reduction.

It is proved that solving minimum reduction of decision table is a NP hard problem [10]. Therefore, heuristic algorithms are used for obtaining suboptimal solutions. Currently, an excellent decision table reduction algorithm includes Genetic Algorithms (GA) based reduction [11], discernibility matrix [12], entropy based dimensionality reduction [13], etc.

Genetic Algorithms based reduction and entropy based reduction have good performance in dimensionality reduction. However, they are not suitable for text mining tasks because the three characteristics which were stated in the end of last subsection. Reduction algorithm based on a discernibility matrix needs to find the attribution kernel of the matrix. Unfortunately, the candidate feature set of a text mining problem is too large and the feature distribution is extremely uneven. In this case, usually the attribute core is an empty set. Therefore, a discernibility matrix is not suitable for text feature reduction.

Improved Quick Reduce (IQR) [14] is based on optimal principle. The optimal principle requires traversing the attributions. However, because of the sparsity and the large scale, blind traversing will lead to huge time and computational overhead. Note that reduction based on the optimal principle could benefit to text classification. If a feature has a strong ability in classification, the ability of it together with other features should also be strong. Therefore, Improved Quick Reduce can be used as a candidate reduction algorithm. The work flow of IQR is shown in Figure 2:

Figure 2. Flowchart of weak classifier weighting

C is the original attribution set, D is the decision set, C' is the preprocessed attribution set, R is the reduction feature set, f is $\delta_{R \sqcup \{x\}}$, h is $\delta_T(D)$, α is $\delta_R(D)$ and β is $\delta_C(D)$.

4.3. Feature Co-Occurrence Quick Reduction

Improved Quick Reduce can achieve feature selection with a lower runtime complexity. This is good, except it ignored the feature co-occurrence in dimensionality reduction. However, features in pairs sometimes play a crucial role in text categorization. For instance, in a set of scientific research papers, *information* and *theory* are not representative features. They are probably discarded when IQR makes attribution reduction because they are too frequent in the document set. The risk of this strategy is the important term *information theory* will be excluded from the final feature set. Due to the existence of this phenomenon, rough set based feature selection methods usually face to “the fastest way but not the best way” problem which limits the precision of classification [15]. To enhance the performance of attribution reduction, we should take feature co-occurrence into account.

Actually, the constraint of attribution reduction in automatic text classification is loose because the reduction is neither necessary to be a minimum reduction nor necessary to be an optimal reduction. Therefore, we can set several rules for reduction to get higher efficiency by sacrificing some completeness while keeping the accuracy in an acceptable degree.

Fully non-necessary conditions for determining whether an attribution is important are:

- (1) The feature has a high overall weight in a particular class.
- (2) The feature's distribution in the class is even relative to other features.
- (3) The feature has low overall weights in other categories.
- (4) The feature has a high weight together with a co-occurrence feature.

The rules stated above construct the foundation of the Feature Co-occurrence Quick Reduction (FCQR) algorithm. The first three conditions can be easily achieved by introducing mathematical expectation and variance into reduction decision-making. To determine whether features are co-occurrence, the covariance and conditional probability judgments should be used. Assuming the occurrence frequency (expectation) of feature f_1 is ζ and the occurrence frequency of feature f_2 is ξ , the correlation of f_1 and f_2 can be calculated as:

$$\rho(f_1, f_2) = \frac{E[(f_1 - \zeta)(f_2 - \xi)]}{\sqrt{E(f_1 - \zeta)^2} \cdot \sqrt{E(f_2 - \xi)^2}} \quad (9)$$

Equation (9) can reflect the degree of contact between two features. However, when a feature has a very high occurrence frequency in the document, it may have a higher connection degree with other features by accident. Coincidental factors can be eliminated by following function:

$$\sigma = \frac{\min(P(f_1|f_2), P(f_2|f_1))}{P(f_1) \cdot P(f_2)} \quad (10)$$

In Equation (10) when σ is lower than the preset threshold, the connection between f_1 and f_2 will be treated as random.

According to the analysis above, the detail of FCQR can be reflected by its Pseudo code as shown in Figure 3.

4.4. Complete Steps of the Dimensionality Reduction Method

Hitherto, the novel rough set based framework for text feature dimensionality reduction algorithm is proposed. The complete process of the novel method is shown in Figure 4.

Using the steps shown in Figure 4, the feature co-occurrence based rough set method for dimensionality reduction is constructed. It can be used for image feature reduction, text feature reduction, structured data feature reduction, etc.

Algorithm: Feature Co-occurrence Quick Reduction (FCQR)

Input: feature set $F = \{f_1, f_2, \dots, f_m\}$ upper bound, lower Bound, threshold T_1 and T_2

Output: selected feature set R

```

1  begin
2  set  $R = \Phi$ 
3  for (  $i=0, i \leq m, i++$ )
4      calculate expectations, variances, covariances, etc.
5      find co-occurred features
6      if (  $f_i > \text{upper bound} \ \& \ f_i$  is not co-occurred feature)
7          discard  $f_i$ 
8      else if ( lower bound  $< f_i < \text{upper bound}$ )
9           $R = R + f_i$ 
10     else if (  $f_i < \text{lower bound} \ \& \ \text{not co-occurred feature}$ )
11         discard  $f_i$ 
12     else if (  $f_i$  is a co-occurred feature)
13          $\rho > T_1 \ \& \ \sigma > T_2 ? R = R + f_i : R$ 
14     else
15         print("error!")
16  for (  $i=0, i \leq m, i++$ )
17      if (  $\delta_{R \cup \{x\}} > \delta_R(D)$  )
18           $R = T$ 
19      else
20          continue
21  end

```

Figure 3. Pseudo code of FCQR

Figure 4. Pseudo code of FCQR

5. Review of Routh Set

We used English texts downloaded from Reuters 21578 and Chinese texts downloaded from Sougo Corpus as the training set and test set. The most intuitive standard of the reduction performance is the difference between the numbers of features as shown in Table 2:

Table 2. Number of features before and after the reduction

Number of documents	Original feature number	Reduction feature number	Number ratio
100	2153	34	63.32
300	3362	47	63.83
1000	4101	67	61.21
3000	4887	74	66.04
10000	5239	83	63.12

As shown in Table 2 the feature size can be reduced more than 60 times by the FCQR based rough set method. The purpose of dimensionality reduction is improving the accuracy and efficiency of text classification. Therefore, long reduction time will lose its meaning of efficiency. We measured the time consumption of the novel method proposed in this paper and compared its runtime overhead with GA based reduction [16], discernibility matrix reduction [17] and entropy based reduction [18] with different size of document sets. Time consumption of different algorithms are shown in Figure 5:

Figure 5. Convergence rate of FullBoost

Figure 5 reveals the novel dimensionality reduction method proposed in this article has a lower runtime complexity than many other classic rough set based reduction algorithms. In order to facilitate the display, the logarithmic scale is used in X-axis.

Moreover, we use experiments to measure the total time consumption of dimensionality reduction and classification. The feature selection is based on the traditional rough set method and FCQR. The classification algorithms include SVM, Naïve Bayes, AdaBoost and C 4.5. Experimental results reveal dimensionality reduction based on FCQR can lead to an average of 23% time saving over different classification algorithms.

As we know, precision of reduction is not easy to be evaluated directly. However, the classification based on the selected feature set can reflect the accuracy of reduction. We randomly downloaded 12000 documents from Reuters 21578. They belonged to six categories: politics, economics, education, sports, arts and science. Each category has 2000 text, 1000 of them represent the training set and the other half represents the test set. We test classification precision using Support Vector Machine (SVM), Naïve Bayes (NB) and Neural Networks (NN) based on the reduced feature set. The result is shown in Table 3:

Table 3. Precision of classification algorithms based on FCQR

Algorithm Class	SVM	Naïve Bayes	Neural Networks
Politics	0.895	0.813	0.864
Economics	0.897	0.799	0.853
Education	0.912	0.830	0.862
Sports	0.909	0.824	0.841
Culture & Art	0.911	0.822	0.867
Science	0.898	0.815	0.855

As shown in Table 3, text categorization based on the feature set selected by the FCQR-rough set has ideal accuracy. The overall classification precision is more than 90 percent, especially when using SVM,

Moreover, we compared the precision of several classification algorithms based on diffident dimensionality reduction methods. The experimental data is shown in Figure 6.

Figure 6. Precision of classification algorithms based on different dimensionality reduction methods

As shown in above figure, the novel feature selection method can improve classification accuracy no matter if the categorization algorithm is Naïve Bayes, SVM or neural networks. The experimental result reveals more representative features can be selected by FCQR and for this reason, the classification accuracy is improved.

Because it is not easy to evaluate accuracy of dimensionality reduction, we carry out a lot of manual annotation and manual comparison to estimate accuracy of FCQR and several control methods. The manual labeling dataset contains 10 thousands of short text and 10 thousand structured data records, the dataset were labeled by many experts in related fields. The evaluation result of feature selection precision of the novel method is shown in Figure 7.

Figure 7. Evaluation result of dimensionality reduction precision

6. Conclusion

A rough set based feature selection method is proposed in this article. In order to reduce the decision table, we designed the FCQR algorithm which takes the feature co-occurrence into account to achieve fast reduction while keeping the accuracy in an excellent degree. In this way, the curse of dimensionality and feature sparsity problems in traditional text classification tasks are avoided. Experimental results reveal that the novel method has a higher accuracy and lower runtime complexity than many vector space model based and probabilistic topic model based reduction algorithms. Furthermore, its performance is better than many previous rough set based dimensionality reduction algorithms.

Fuzzy boundary is an emerging tool in rough set theory. Using it may further improve the performance of FCQR based rough set reduction. This will be undertaken as future work on this topic.

Acknowledgement

This work was supported partly by “the Fundamental Research Funds for the Central Universities ” in UIBE no. 15QD07.

References

1. A. Yindalon, D. F. Lawrence, Z. Li, R. P. Eric, and E. Efstathiadis, “A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization,” *Journal of the Association for Information Science and Technology*, Vol. 42, No. 10, pp. 1964-1987, October 2014
2. P. T. Fernando, J. Cardiff, P. Rosso, and P. David, “Weblog and Short Text Feature Extraction and Impact on Categorization,” *Journal of Intelligent & Fuzzy Systems*, Vol. 27, No. 52014, pp. 2529-2544, May 2016
3. R. S. Bhalerao, J. Y. Ollitrault, and P. Subrata, “Principal Component Analysis of Event-by-Event Fluctuations,” *Physical Review Letters*, Vol. 114, pp. 1-6, December 2015
4. M. Z. Ming, “Sparse Principal Component Analysis and Iterative Thresholding,” *Annals of Statistics*, Vol. 41, No. 2, pp. 772-801, February 2013
5. J. Guia, Z. Suna, and W. Jia, “Discriminant Sparse Neighborhood Preserving Embedding for Face Recognition,” *Pattern Recognition*, Vol. 45, No. 8, pp. 2884-2893, August 2012
6. R. Karbauskaitė, O. Kurasova, and G. Dzemyda, “Selection of the Number of Neighbors of Each Data Point for the Locally Linear Embedding Algorithm,” *Information Technology and Control*, Vol. 36, No. 4, pp. 359-364, April 2015
7. F. Y. Cao, J. Y. Liang, and D. Y. Li, “A Dissimilarity Measure for the k-Modes Clustering Algorithm,” *KNOWLEDGE-BASED SYSTEMS*, Vol. 26, No. 2, pp. 120-127, May 2014
8. K. Devarajan, G. L. Wang, and N. Ebrahimi, “A Unified Statistical Approach to Non-Negative Matrix Factorization and Probabilistic Latent Semantic Indexing,” *MACHINE LEARNING*, Vol. 99, pp. 137-163, 2015
9. D. Q. Miao, H. Y. Duan, and N. J. Zhang, “Rough Set based Hybrid Algorithm for Text Classification,” *Expert Systems with Applications*, Vol. 36, No. 5, pp. 9168-9174, 2009
10. Y. S. Lee, L. Rocky, C. Y. Chen, P. C. Lin, and J. C. Wang, “News Topics Categorization using Latent Dirichlet Allocation and Sparse Representation Classifier,” in *Proceedings of IEEE International Conference on Consumer Electronics*, pp. 126-13, Taipei, Taiwan, June 2015
11. Polkowski, Lech, Shusaku Tsumoto, and Tsau Y. Lin, *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, Vol. 56, Physica, 2012
12. J. Dai and H. Tian, “Fuzzy Rough Set Model for Set-Valued Data,” *Fuzzy Set and Systems*, Vol. 229, No. 7, pp. 54-68, July 2013
13. S. B. Kanti, S. S. Sankar, and C. Kripasindhu, “A Genetic Algorithm-based Rule Extraction System,” *Applied Soft Computing*, Vol. 12, No. 1, pp. 238-254, January 2013
14. C. C. Yeh, D. J. Chi, T. Y. Lin, and S. H. Chiu, “A Hybrid Detecting Fraudulent Financial Statements Model using Rough Set Theory and Support Vector Machines,” *Cybernetics and Systems*, Vol. 47, No. 4, pp. 261-276, April 2016
15. H. M. Chen, T. R. Li, C. Luo, S. J. Horng, and G. Y. Wang, “A Decision-Theoretic Rough Set Approach for Dynamic Data Mining,” *IEEE Transactions on Fuzzy Systems*, Vol. 23, No. 6, pp. 1958-1970, June 2015
16. X. Y. Chen and S. Y. Wu, “The Optimization Assignment Model of Multi-Sensor Resource Management based on Rough Entropy,” *International Journal of Grid and Utility Computing*, Vol. 8, No. 5, pp. 233-244, August 2015
17. Y. H. Han, Y. Yang, Z. G. Ma, Y. Yan, N. Sebe, and X. F. Zhou, “Semi-Supervised Feature Selection via Spline Regression for Video Semantic Recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 2, pp. 252-264, February 2015
18. W. Z. Wu, Y. H. Xu, M. W. Shao, and G. Y. Wang, “Axiomatic Characterizations of (S, T)-Fuzzy Rough Approximation Operators,” *Information Sciences*, Vol. 334, pp. 17-43, September 2016

Lei La is an associate professor of School of Information Technology & Management, University of International Business and Economics. His research interests include semi-supervised machine learning and network data analysis.

Qimin Cao is an associate research fellow of Library, China University of Political Science and Law. Her research interests include system engineering and information analysis.

Ning Xu is an undergraduate student of School of Information Technology & Management, University of International Business and Economics. His research interests include machine learning algorithm development and coding.