A Text Classification Algorithm Based On RS

Jianlin Li[1, 2]

[1]Department of Computing & Software, Nanjing College of Information Technology, Jiangsu, 210023, China

[2]Department of Computer Science, University of Regina, Saskatchewan, S4S0A2, Canada

email: lijl@njcit.cn

Keywords: Rough set, Meta-feature selection, Attribute reduction, Text classification

Abstract. Study a variety of text feature extraction methods, through mutual information(MI), document frequency(DF),information gain(IG) and χ2 statistics(CHI) algorithm, using of their respective advantages complementary, proposed a kind of multiple combination feature extraction algorithm based on rough set(RS-MCFEA);First using attribute reduction based on rough set in keeping the classification ability under the condition of constant fast will text feature space dimension reduction, and then by multiple combinations of features extracted in the feature space after the dimension reduction is more representative of characteristic items, filter out some representative weak feature items, finally using SVM classifier to classify text; The experimental results show that this algorithm can effectively improve text classification accuracy and efficiency of classification.

Introduction

At present, the text feature extraction is mainly on the basis of the characteristics of document matrix by evaluation each feature through evaluation function, by setting threshold reserve a certain number of features, it is often lead to the loss of useful information [1-4];This paper propose a kind of multiple combination feature extraction algorithm based on RS(RS-MCFEA), through the rough set attribute reduction quickly get a reduction, after the original feature matrix dimension reduction, use the combination method to overcome the defect of a single feature extraction, and finally obtain better classification characteristics, the experimental results show that the RS-MCFEA extraction algorithm to improve the characteristics of words text classification accuracy and efficiency are effective.

Rough Set Theory. Knowledge reduction is the kernel of the rough set theory, it is under the condition of knowledge classification ability unchanged, to delete irrelevant or non-important attributes and attribute values of the process.

For an information system $S=(U,A,V,f)$. U is a finite nonempty set of object, $U=\{x_1,x_2,...,x_n\}$, $A=C\bigcup D$ as attributes of the nonempty finite sets, $C\bigcap D=\varnothing$, subsets C and D are known as condition attributes set and decision attribute set. $V=\bigcup_{a\in A}V_a$ is attribute value set, V_a is an nonempty set of value of $a\in A$, $f:U\times A\to V$ is an information function that maps an object in U to exactly one value in V_a.

Definition 1: For a attributes subset $P\subseteq A$, its information entropy $H(P)$ is defined by [4]:

$$H(P)=-\sum_{i=1}^{m}P(X_i)\log P(X_i). \tag{1}$$

Where $U/IND(P)=\{X_1,X_2,...,X_n\}$, $P(X_i)=\frac{|X_i|}{|U|}$ $(i=1,2,...,m)$.

Definition 2: For an information system $S=(U,A,V,f)$, the importance of attribute $a\in A$ is defined by:

$$Sig_A(a)=|H(A)-H(A-\{a\})|. \tag{2}$$

Definition 3: For any nonempty set of $X \subseteq A$, the super-core of X defined by:

$$Core_S(X) = \{x \in X \mid Sig_A(x) > 0\}. \quad (3)$$

The attribute importance of $a \in A$ in A is by the change of information entropy to measure caused by the removing a from A. If $Sig_A(a) > 0$, says $a \in A$ in A is necessary, a is optional indicator, if $Sig_A(a) = 0$, then a is redundant, should be deleted from A [4].

Definition 4: For a decision tables $S = (U, C \bigcup D, V, f)$, C as condition attribute set, D as decision attribute set, $\varnothing \neq P \subseteq C$, D about P positive domain $POS_P(D)$ is defined by: $POS_P(D) = \bigcup_{Y_i \in U/D} \underline{P}(Y_i)$, $\underline{P}(Y_i) = \bigcup_j \{Z_j \mid Z_j \in U/P \text{且} Z_j \in Y_i\}$ is lower approximation set of Y_i, $POS_P(D)$ is according to the classification of U/P information can be accurately divided to the decision attribute D equivalence class to a set of objects.

Definition 5: For a decision tables $S = (U, C \bigcup D, V, f)$, C as condition attributes, D as decision attribute set, $\varnothing \neq P \subseteq C$, $a \in Q$, if $POS_{Q-\{a\}}(D) = POS_Q(D)$, a in Q can be omitted, otherwise a known as not be omitted in Q. If each a in Q is not be omitted and $POS_Q(D) = POS_C(D)$, we call Q is D-reduct of C.

Feature Reduction Algorithm

Algorithm 1 Given a non-empty set $X \subseteq A$, calculate the generalized nuclear of X.

Input: attribute subset $X = \{x_1, x_2, \ldots, x_k\}$.

Output: the generalized nuclear of X :Core (X).

(1) Core(X)←Φ;

(2) calculate H (X);

(3) i←1;

(4) calculate $H(X-\{x_i\})$

(5)IF ($Sig(x_i)=|H(X)–H(X-\{x_i\})|)>0$ Then Core(X)←Core(X) ∪ $\{x_i\}$;

(6) IF $i \geq k$, turn (7),Else i←i+1,turn (4);

(7)output Core(X);

The time complexity of algorithm 1 is $TC \sim O(X^2 U^2)$

Algorithm 2 Calculate a reduction of $X \subseteq A$.

Input: the generalized nuclear of X: Core (X).

Output: a reduction of X.

(1) Reduct←Core(X);

(2)For $(i = |\mathrm{Re}\,duct|, 1 \leq i \leq |\mathrm{Re}\,duct|, i--)$ If $POS_{\mathrm{Re}\,duct - x_i}(D) \neq POS_C(D)$ Then $\mathrm{Re}\,duct = \mathrm{Re}\,duct$ Else $\mathrm{Re}\,duct = \mathrm{Re}\,duct - x_i$, End If, End For;

(3)output Reduct, finished;

The time complexity of algorithm 2 is $TC \sim O(X^3 U^2)$

RS-MCFEA

Algorithm implementation process as shown in figure 1:

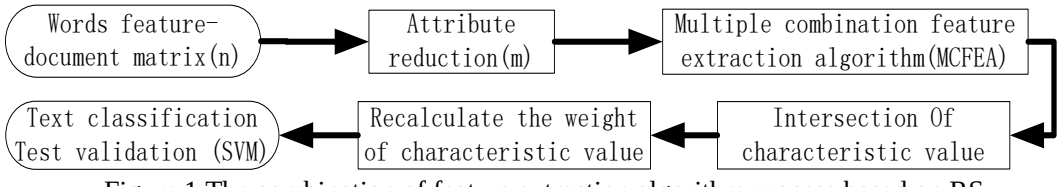

Figure 1 The combination of feature extraction algorithm process based on RS

(1) Prepare the data; The data is divided into training set and test set. Assume that the training set of document number for N, set to x_1, x_2,…, x_N. We select n keywords a_1, a_2, …, a_n as document feature, can get a word feature-document matrix A, a row of the matrix represents a document feature vector, a list of matrix represents the frequency of the appear of a feature word in the document .

(2) Construct decision table, according to the classification of the information in advance, we can construct decision table D_T.

(3) Use algorithm 1, 2 to attribute reduction for decision table，Get dimension reduction of decision table D_T' ,it can retain all the feature information of decision table D_T.

(4)Multiple combination feature extraction: using MI DF CHI and IG calculate the weight of each word in the decision table D_T' after dimensionality reduction; Each method calculate weights of the every words and sorted by weight, obtain four array；

1)Taken T+t feature words from front of the weight array, and calculate the intersection of four arrays, the number of intersection is K.

2)if K=T, turn 3); if K<T, t=t+1 turn 1); if K>T, t=t-1 turn 1).

3)finish.

(5) Use the DF re-evaluated the weight of the rest of the T feature words, this weight values as the final weight of the feature words that composition of this word feature-document matrix (N×T matrix) of the training sample.

(6) By the above combination of feature extraction algorithm, feature items and corresponding data take into the SVM classifier for classification training and testing.

Experimental Results and Analysis

We use the corpus which provides by Chinese academy of sciences institute of computing the natural language processing open platform to experiment, select the politics economy military culture industry and computer six categories total 2000 articles. We use the proposed RS-MCFEA feature extraction methods select 1500 articles as training text, the rest of the 500 articles as a test text validation.

We used the IG MI DF CHI method and the RS-MCFEA feature extraction then using SVM classifier after training and classification accuracy are compared, and the experimental results are as follows Table 1 and Table 2 below:

Table 1 popular single feature extraction method+SVM classification results

Feature extraction method	Feature extraction method +SVM			Classification accuracy (%)
	Train time(S)	Test time(S)	total time(S)	
MI	149	31	180	63.5
DF	1048	92	1140	87.3
CHI	405	48	453	85.2
IG	1392	159	1551	86.1

Table 2 Feature extraction based on RS-MCFEA+SVM classification results

Feature reduction time(S)	MCFEA time(S)	SVM		Total time(S)	Classification accuracy (%)
		Train time(S)	Test time(S)		
65	43	43	27	**178**	**91.6**

The total time of using RS-MCFEA is 178 seconds, lower than the currently popular feature extraction based on threshold method using SVM classifier for training and classification of the time, at the same time, the classification accuracy is 91.6%, also has a substantial increase.

Conclusion

At present, although there are many more classic text classification algorithm, some scholars also

improved and optimization for these classical algorithms [1-4], but most of the classification system is through the setting threshold keep a certain number of features to complete, this makes the system classification algorithm performance is congenital deficiency restriction. Attribute reduction is the important topic of rough set research, many scholars in this respect studied, but now have proved that it is a typical NP-hard problem [5,6,7], and that is when the number of objects or attributes of the decision table is very large, the current reduction algorithm to calculate minimum reduction is still very difficult. And this paper is to use the generalized nuclear and the definition of reduction to get a relative reduction of the feature matrix, not pursuing minimum or optimal reduction, and finally to through the multiple combination feature extraction algorithm get the best classification feature subset and decision rules, the results show that this method can greatly improve the system operation speed and get the higher classification accuracy.

Acknowledgement

In this paper, the research was sponsored by the year 2016 "Blue Project" of Jiangsu Province young academic leaders (Project No. Jiangsu Education 2016-15).

References

[1] Shen Gong. Text classification feature extraction methods and improved [J], computer simulation, (3): 222-224(2006)

[2] Kang Tao. Text feature extraction method based on PCA and RS[J], Computer applications, (10):88-90(2007)

[3] Yang Chuang, Yang Bingru. The combined extraction characteristics of text classification technology based on rough set[J]. Computer application research, (7):97-99(2007)

[4] Liu Jian, Zhang Weiming. Mutual information text feature selection method and improved [J], computer engineering and application,44 (10):135-137(2008)

[5] Y.Y. Yao, Y. Zhao, Discernibility matrix simplication for constructing attribute reducts, Information Sciences, Vol. 179, No. 5, 867-882,(2009)

[6] J.T. Yao, Y.Y. Yao. Induction of Classification Rules by Granular Computing[J], Rough Sets and Current Trends in Computing,950-950(2002)

[7] J.L. Li. An approach to meta feature selection, CCECE 2013, pp.210-214,(2013).

[8] J.L. Li, X.F. Deng, Y.Y. Yao. Multistage Email Spam Filtering Based on Three-Way Decisions, RSKT 2013, LNAI8171, pp.313-324,(2013)