A complementary hybrid classification algorithm based on Web text

Lili Xing*, Bing Zhang, Yuhong Lu, Zhong Li

Institute of Disaster Prevention, Information Department, Yanjiao 065201, China

Received 17 September 2014, www.cmnt.lv

Abstract

In view of the insufficiency of existing weight computation methods and SVM algorithm, a weight computation method of variable precision rough set based on Web text and a complementary hybrid classification algorithm are proposed; In the hybrid classification algorithm, the rough set is used as a front-end processor of SVM, the traditional SVM is optimized from classification efficiency and precision through the reduction theory and weight computation method proposed in this paper. The experimental results show that the reduced and weighted data are classified using SVM, and then the performance of classification is further guaranteed.

Keywords: SVM, rough set, reduction, weighting, web text classification

1 Introduction

Support vector machine (SVM) is based on the VC dimension theory and structural risk minimization principle in the statistical theory, to find the best compromise between model complexity and learning ability according to the limited sample information, in order to obtain the best generalization ability. SVM not only can solve some practical problems in many learning methods, small sample, over learning, high dimension, and local minimum, but also has strong generalization ability, then has the great application potential in the Web text classification field.

The classical SVM algorithm transfers the classification problems into quadratic programming, realizes the optimization of the classification hyper plane, while the calculation amount of quadratic programming increases exponentially with the increase of variables. When there are many sample attributes, SVM has complex network structure and slow recognition speed. Although the dimension reduction has been conducted before classification through feature selection before classification, redundant features inevitably exist in the selected feature sub due to a lot of synonyms and polysemants in natural languages, which limits the application of SVM algorithm, especially for pattern classification of large data. It's an important issue how to further improve the SVM model, training algorithm, and the real-time of data processing based on SVM, shorten training time, and increase the classification accuracy.

2 Rough set theory

2.1 KNOWLEDGE REPRESENTATION SYSTEM

In rough set theory, knowledge is considered to be an ability classifying the realistic or abstract object. A knowledge representation system can be expressed as a quadruple $S =< U, A, V, f >$, in which U means all the object, and is called as domain (nonempty finite set), A is a nonempty finite set of U, and is called as attribute set, V is attribute value set, $V = \bigcup_{a \in A} V_a$, V_a is the attribute value range of attribute $a \in A$, f is a mapping relationship, $f : U \times A \to V$ is a information function, and means $f(x, a) \in V$ when $\forall x \in U$ and $a \in A$. If A includes the condition attribute set C and decision attribute set D, C and D satisfy $A = C \cup D$ and $C \bigcap D = \Phi$, then S is called as decision table.

2.2 REDUCTION

Set U to be the domain and R be the indistinguishable relationship in U. If $ind(R) = ind(R - \{r\})$, and $r \in R$, then r can be defined as to the able-to-be-reduced knowledge (relationship), and then R is correlative (having the able-to-be-reduced property); otherwise, if any $r \in R$ is unable to be reduced, then R is defined as independent. Set $P \subseteq R$, and $ind(P) = ind(R)$, if P is independent, P is set as a reduction of R.

2.3 VARIABLE PRECISION ROUGH SET MODEL

In the variable precision rough set (VPRS) model, the approximation containing threshold is set to relax the strict boundary definition of the classical rough set, then the model has a certain ability to overcome the noise. The lower and upper approximation set of Variable precision α can be defined as shown in Equations (1) and (2):

* ***Corresponding author's*** e-mail: Xinglili@cidp.edu.cn

$$\underline{R_\alpha}(X)=\{x\in U: C([x]_R, X)\geq \alpha\}, \quad (1)$$

$$\overline{R_\alpha}(X)=\{x\in U: C([x]_R, X)>1-\alpha\}, \quad (2)$$

$$C([x]_R, X)=\begin{cases}|[x]_R\cap X|/|[x]_R|, |X|>0\\ 0, |X|=0\end{cases}. \quad (3)$$

In Equation (3), the inclusion degree $C([x]_R, X)$ means the degree in which the equivalence class of x formed by the relationship R is contained by the set X, and $\alpha\in(0.5,1]$ is the user-specified threshold of inclusion degree. The equivalence classes in the relationship R are included in X in the inclusion degree larger than α, and the elements of these equivalence classes form the lower approximation of α. The upper approximation of α is the element set of the equivalence class in U included in X in the inclusion degree larger than $1-\alpha$. When $\alpha=1$, the model degrades into a standard rough set model.

3 Weight computation method of variable precision rough set based on web text (WVPRS)

The widely used inverse document frequency (IDF) weighting method only represents the distribution important degree of feature in the whole sample from the overall classification, but does not take into account the characteristics of the Web text and the existing classification decision information; in view of this, this paper constructs a new method feature weighting computation method based on Web text.

3.1 FEATURE OF WEB TEXT AND WEITHTING SCHEME OF HTML TEXT TAGS

There are a lot of HTML format texts in the Web text, compared with the ordinary text, there are obvious identifiers in the HTML text, and more obvious structure information.

Through analyzing the format of HTML file, consider the following markups:

1) Title <TITLE>: TITLE summarizes the entire content of the webpage, so it plays a key role in classification.

2) All levels of headings <H1>, <H2>,…, <H6>: the contents of H1, H2,…, H6 specifically describes the basic structure of the webpage, the important degree is from H1 to H6 decreased.

3) Bold <B>，Underline <U>，Italic <I>：B, U, and I change display effect of the text, play an emphasis, reflect the content correlation from a certain side.

4) The data in Meta also provide some useful information, but its format is not standardized and often does not appear, in this paper this part is not considered.

Then we can acquire Hyper Text Markup set S={TITLE, Hl , H2, H3, H4,H5, H6, B, U, I}, the weighting scheme of HTML markup is designed in this paper (shown in Table 1).

TABLE 1 Weighting scheme of HTML markup

HTML markup θ	Weight coefficient λ
<TITLE>	7
<H1>	5
<H2>	4
<H3>	3
<H4>	2
<H5>	2
<H6>	2
<B>	3
<U>	3
<I>	3

3.2 WEIGHTING COMPUTATION OF WVPRS

According to variable precision rough set model, n types of sample set are divided into the equivalence class of $U=\{D_1, D_2, ..., D_n\}$ based on the class number of the text. R is the equivalence relationship by different feature words. α - lower, upper approximation set, can be defined as shown in Equations (4) and (5):

$$\underline{R_\alpha}U=\{\underline{R_\alpha}D_1, \underline{R_\alpha}D_2, ..., \underline{R_\alpha}D_n\}, \quad (4)$$

$$\overline{R_\alpha}U=\{\overline{R_\alpha}D_1, \overline{R_\alpha}D_2, ..., \overline{R_\alpha}D_n\}, \quad (5)$$

where $\underline{R_\alpha}D_i$, $\overline{R_\alpha}D_i$ respectively represent the lower, upper approximation set of α under the equivalence relationship in the set D_i . $\underline{R_\alpha}U$, $\overline{R_\alpha}U$ respectively describe the element set of the equivalence class that belongs to the equivalence class of the equivalence relationship R for division U and must or may be included in decision attribute division. In the literature 8, the approximation quality of classification $\gamma_R(U)$ (shown in Equation (6)) was introduced to measure the accuracy to this division, and was used to weight feature words frequency instead of the inverse document frequency, and the experimental results showed that this method can effectively improve divisibility of the sample.

$$\gamma_R(U)=\frac{\sum_{k=1}^{n}\left|\underline{R_\alpha}D_k\right|}{|U|}. \quad (6)$$

According to the above analysis, combined with the characteristics of Web text, a weighting computation method Variable Precision Rough Set based on Web text (WVPRS) is put forward to double weight the feature frequencies, as shown in Equation (7):

$$Hw_{ij}=\lambda_\theta\times tf_{ij}\times\frac{\sum_{k=1}^{n}\left|\underline{R_a}D_k\right|}{|U|}, \quad (7)$$

where, λ_θ means the weighted coefficients corresponding to webpage markup. Where $\theta \in S$, S={TITLE, H1, H2, H3, H4, H5, H6, B, U, I}, as shown in Table 1. tf_{ij} is the frequency of the j feature word in the i^{th} text; n is the number of the classes in the sample set. $\underline{R}_\alpha D_k$ is the lower approximate set of α of set D_k under the relationship R, where n sample set is divided into the equivalence class $U = \{D_1, D_2, ..., D_n\}$ of according to the type number of text.

In Equation (7), on the one hand, the weight computation method considers the effects of Web text special markup, appropriately weights on special markups, so as to improve the importance of the low frequency feature with high classification ability; on the other hand, this method introduces the classification decision information into the weight formula, calculates the degree of consistency between the division of feature word and all kinds of decision classification, and then sums the degree of consistency between the feature words and the overall decision-making, to reflect the overall weight from the important degree of the feature words to all kinds of classification, entirely representing the important information of feature words for all the classification. The WVPRS method more accurately describes the important degree of every feature word for classification in the Web text classification in theory, which is verified in Section 5.4.

4 SVM classification algorithm based on rough set and weight

In this paper, the rough set theory is combined with SVM to put forward a complementary hybrid classification algorithm based on Web text.

The model is composed of the front processing terminal based on rough set and the SVM based back classification. Firstly, the Web text set is pre-treated. After the feature is extracted and selected, a decision table is constructed for feature set, and rough set is used as the front terminal processing tool. Without any loss of effective information, rough set reduction is used for reduction of decision table to remove redundant attributes in the decision table and delete conflict object. After the rough set reduction, the input amount of support vector machines will be greatly reduced, accordingly, the complexity of the support vector machine classification is reduced, and thereby training time is saved, and over-fitting of the training model is avoided in different degrees, but the classification performance is not reduced. Then the WVPRS method does weight computation for the retained features to get intermediate representation of the text. Finally, the algorithm is transferred from the front to the back, and the reduced and weighted data are classified based on SVM, further improving the classification accuracy, reducing calculated amount, shortening training time, and improving classification efficiency.

In the training part of the model, feature subset is obtained after feature selection, a decision table is constructed and reduced, and the weights are computed for the retained features, finally, the SVM is used for training to obtain a classifier; and in the test part, for the text to classify, the feature is extracted and the weight is computed based on the reduction results during the training part, which is finally input to the classifier to do classification, and get classification results.

5 Experiments

In this paper, the software, Webdup 0.93 Beta, is used to obtain the Web Chinese text training set and the test set; ROSETTA is for computation of rough set theory; LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) developed by Zhiren Lin (Chih-Jen Lin) at Department of Information, Nation Taiwan University is adopted as the SVM model. The hardware environment is PC, the CPU is Intel Celeron 1.70GHz, and the memory is 256M.

5.1 ACQUISITION AND PREPROCESSING OF WEB TEXT

In Web text training set, Htrain and Web text test set, Htest, each set includes 10 categories, totally 3432 text. The 10 categories in each website grab have been perfectly classified. Data set distribution is shown in Table 2.

TABLE 2 Experimental data set distribution

No.	Class name	htrain	htest
1	Computer	182	110
2	Art	120	85
3	Education	235	122
4	Traffic	140	76
5	Environment	268	106
6	Economics	300	158
7	Medicine	252	137
8	Military	186	101
9	Politics	329	153
10	Sports	246	126
	Sum	2258	1174

HTML document source codes are scanned; the contents modified markups that need to be weighted are weighted, while those that do not need to be weighted are given 1 as weights according to the default weight value. The word segmentation software, ICTCLAS developed by Institute of software research, Chinese Academy of Sciences, is used to do word segment and some preprocess, such as to stop words.

5.2 FEATURE SELECTION AND CONSTRUCTING DECISION TABLE

The main idea to construct decision table proposed in the literature [9] is to compute weights of features in the text and then to obtain a decision table in real type. Because the

rough set can only do attribute reduction for decision table in discrete type, the decision table must be discretized next. The disadvantage of this approach is that, the quality of discretization for real-valued decision table will directly affect the quality of text classification.

In this paper, the method to construct decision table is based on feature selection. Since the retained features after selection are more important features for classification, then it does not affect the classification quality to determine attribute value according to the simple structure of Web text, and a smaller discrete decision table is obtained at the same time, which reduces the workload of decision table reduction.

First, 1000 features are retained by the information gain feature selection method, and then the attribute values are determined through testing whether these features appear in the Web text and their locations, to directly obtain the discrete decision table. The method is described as follows: the set of all text in each class is as the domain and each text in the domain is as the objects; text category is as decision attribute and a text category is decision attribute value; the retained feature set after feature selection is conditional attribute set, namely each feature is condition attribute; the following method is to determine the condition attributes value of features in the object: if the feature does not appear in the text, the value is 0; otherwise, if the feature occurs only in the text, the value is 1; if it appears on the markup, then the value refers to Table 1, the attribute value is the maximum weight coefficient of its markup. For example, if a certain feature appears in both <TITLE> and <H2>, the condition attribute value in the object is 7.The decision table is constructed, shown in Table 3.

TABLE 3 Classification decision table based on Web text

Text	Feature					Class
	Football	Profit	Floppy disk	Software	Examination	
D1	7	0	0	0	0	10
D2	1	5	0	0	0	6
D3	0	0	5	7	0	1
…						
Dn	0	0	0	7	5	3

5.3 REDUCTION OF DECISION TABLE

In this experiment, the attribute reduction algorithm with identification matrix and logic operation is used to do attribute reduction, after which, 657 features are retained, and the rate of "attribute evaporation" is 34.3%. It can be seen that the reduction algorithm based on rough set is an effective tool for "data concentrated".

5.4 WVPRS WEIGHT COMPUTATION AND SVM CLASSIFICATION

The weights of the retained features after reduction are computed using the WVPRS weight computation method proposed in this paper (set $\alpha = 0.8$), finally the text vector is obtained, the classifier is trained by SVM, and this classifier is compared with the SVM classifier before reduction or the SVM classifier by other weight computation methods.

In the experiment, the SVM classifier with linear kernel function is used; the precision ratio, recall ratio and F1 value are used as indicators to compare and analyzed the classification results of different methods.

The experimental result in Table 4 shows the performance of classifier before reduction by the weight computation $Hw_{ij} = tf_{ij} \times idf_j$ and $Hw_{ij} = \lambda_\theta \times tf_{ij} \times idf_j$.

TABLE 4 Performance of classifier before reduction

Name	$Hw_{ij} = tf_{ij} \times idf_j$			$Hw_{ij} = \lambda_\theta \times tf_{ij} \times idf_j$		
	Precision ratio %	Recall ratio %	F1 %	Precision ratio %	Recall ratio %	F1 %
Computer	92.46	90.70	91.57	94.32	93.98	94.15
Art	69.20	20.68	31.84	76.72	25.83	38.65
Education	90.37	91.22	90.79	93.15	91.00	92.06
Traffic	91.89	92.97	92.43	90.86	94.15	92.48
Environment	89.34	91.67	90.49	92.68	92.03	92.35
Economics	81.35	85.36	83.31	83.92	90.06	86.88
Medicine	80.26	86.92	83.46	85.88	90.49	88.12
Military	89.34	88.73	89.03	92.30	91.87	92.08
Politics	73.59	65.36	69.23	75.12	70.23	72.59
Sports	93.00	91.82	92.41	95.03	93.26	94.14
Average performance	85.08	80.54	81.46	88.00	83.29	84.35

From the experimental results in Table 4, it can be seen that for the classifier after weighting markup with the same data set and classification algorithms, the precision ratio is increased by 3.43%, the recall ration is enhanced by 3.12%, F1 is increased by 3.55%, and the average performance is better than that of the classifier of which markups are not weighted. Thus, it can improve the performance of the classifier that the HTML text markups are appropriately weighted.

The experimental result in Table 5 shows the performance of classifier after reduction by the weight computation $Hw_{ij} = \lambda_\theta \times tf_{ij} \times idf_j$ and Equation (7).

TABLE 5 Performance of classifier after reduction

Name	$Hw_{ij} = \lambda_\theta \times tf_{ij} \times idf_j$			$Hw_{ij} = \lambda_\theta \times tf_{ij} \times \frac{\sum_{k=1}^{n} \lvert R_a D_k \rvert}{\lvert U \rvert}$		
	Precision ratio %	Recall ratio %	F1 %	Precision ratio %	Recall ratio %	F1 %
Computer	95.37	93.26	94.30	98.69	94.38	96.49
Art	75.35	28.43	41.28	77.34	32.59	45.86
Education	92.32	92.96	92.64	95.85	94.31	95.07
Traffic	90.59	94.57	92.54	91.46	94.60	93.00
Environment	93.19	93.67	93.43	94.93	95.89	95.41
Economics	84.39	90.32	87.25	88.26	90.13	89.19
Medicine	88.90	91.48	90.17	90.00	90.87	90.43
Military	94.26	91.09	92.65	96.75	94.68	95.70
Politics	76.16	71.49	73.75	82.61	76.30	79.33
Sports	97.00	94.02	95.49	98.85	96.27	97.54
Average performance	88.75	84.13	85.35	91.47	86.00	87.80

From the results in Table 4 and Table 5, it can be seen that after reduction, all kinds of precision ratios and recall ratios are increased or decreased, but all the F1 are improved, the overall average value is also slightly increased; while the experimental results show that, the training time of the classifier is shortened after reduction. The training time of classifier is 160s before reduction, while it is 51s after reduction. Thus, the reduction theory of rough set guarantees the classification performance and effectively simplifies the training set, which greatly reduced the input number of SVM, thus improved the training speed.

According to the comparison between the experimental data in Table 5, it can be analyzed and concluded that, the WVPRS weight computation method based on rough set, proposed in this paper, on the one hand considers the role of markups in Web text; on the other hand, it replaces the inverse frequency using the important degree of each attribute for decision attribute, double weights the frequency of features, which further improves the accuracy of SVM classification, with precision ratio increased by 3.06%, recall ratio enhanced by 2.22%, and F1 value improved by 2.87%.

5.5 EXPERIMENTAL CONCLUSION

The hybrid classification algorithm combines the advantages of rough sets with that of SVM, fosters strengths and circumvents weaknesses. It has the following advantages:

1) On the premise of guaranteeing the classification performance using rough set reduction theory, the redundant features are cancelled out, feature dimension is reduced, which greatly decreased the input data of SVM, and improved the training speed.

2) WVPRS weighting method takes into account the structure characteristics of the Web text, at the same time, introduces the decision information into the importance of features, double weights the frequency of features, more fully characterizing the importance of feature in the Web text for classification, which further improved the post classification accuracy.

3) SVM as a post classifier has the good ability to suppress the noise and perfect generalization performance, which guarantees the classification performance of this method.

6 Conclusions

In this paper, first, the advantages and disadvantages of the SVM for Web text classification were analyzed; then the basic theory of rough set was expounded; according to the deficiency of existing weight computation method, based on the analysis on the characteristics of Web text, the modification role of HTML markers on the webpage content was studied, the weighting strategy of HTML markups was designed, a variable-accuracy rough set weight computation method based on Web text and a hybrid classification algorithm were proposed; in the algorithm, the rough set is used as a front-end processor of SVM, the reduction theory and the weight computation method proposed in this paper respectively optimized the SVM from two aspects: the classification efficiency and classification accuracy; SVM is used as a back-end classifier, the reduced and weighted data were classified using the advantage of SVM, which further guaranteed the classification performance; finally the validity of this algorithm was verified by experiment.

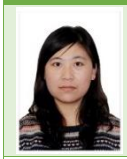
Acknowledgments

This work was supported by the Special Fund of Fundamental Scientific Research Business Expense for Higher School of Central Government (Projects for young teachers) (No.:ZY20130208) and Teachers' Scientific Research Fund of China Earthquake Administration (No.:20110113).

References

[1] Lin Mu 2011 The performance comparison of support vector machine algorithm and other algorithms in text classification based on support *Journal of Inner Mongolia University (Natural Science Edition)* **42**(6) 703-7
[2] Jiang C, Zhang G, Li Z 2011 Anomaly intrusion detection of embedded network system based on SVM optimized by genetic algorithm *Computer applications and software* **28**(2) 287-9
[3] Hu Z, Wang H, Zhang H 2013 Fast support vector machine classification algorithm based on distance sorting *Computer applications and software* **30**(4) 85-7
[4] Ye M, Wu X, Hu X, Hu D 2013 Anonymizing classification data using rough set theory *Knowledge-Based Systems* **43** 82-94
[5] Zhao D, Wang L, Zhang F 2012 Rough set theory of Genetic algorithm applied in dimension reduction of text *Computer engineering and application* **48**(36) 125-8
[6] Fan T-F, Liau C-J ,Liu D-R 2013 Variable consistency and variable precision models for dominance-based fuzzy rough set analysis of possibilistic information systems *International Journal of General Systems* **42**(6) 659-86
[7] Lan J, Shi H, etc. 2011 The related webpage classification based on compound weight of features. *Computer science* **38**(3) 187-90
[8] Hu Q, Xie Z, Yu D 2005 Text classification method based on rough set weight *Information science* **24**(1) 59-63
[9] Yang S 2007 Research on rough set applied in the text classification system *Shandong Normal University*
[10] Luo Q 2003 A RS-Based Application Research on Web Text Mining *Guangxi University*

Authors



Lili Xing, 2/10/1983, Tangshan City, Hebei Province, China.

Current position, grades: full-time teachers, lecturer from 2008 in China Institute of Disaster Prevention.
University studies: Computer Application
Scientific interest: data mining
Publications: 6 papers.

Bing Zhang, 11/4/1983, Qufu City, Shandong Province, China.

Current position, grades: full-time teachers, lecturer from 2008 in China Institute of Disaster Prevention.
University studies: Computer Application
Scientific interest: internet of things
Publications: 6 papers.

Lu yu hong, 1/21/1977, Tangshan City, Hebei Province, China.

Current position, grades: full-time teacher, associate professor from 2012 in China Institute of Disaster Prevention.
University studies: University of Science & Technology Beijing.
Scientific interest: the teaching and research of computer.
Publications: 10 papers.

Li Zhong, 09/05/1966, Zhucheng City, Shandong Province, China.

Current position, grades: Professor, PhD.
University studies: Institute of Disaster Prevention, China.
Scientific interest: artificial intelligence, data mining.
Publications: 60 papers.