A Novel Feature Selection and Attribute Reduction Based on Hybrid IG-RS Approach

Leena H. Patil and Mohammed Atique

Department of Computer Science and Engineering,
Sant Gadge Baba Amravati University, Amravati, India
harshleena23@rediffmail.com,
mohd.atique@gmail.com

Abstract. Document preprocessing and Feature selection are the major problem in the field of data mining, machine learning and pattern recognition. Feature Subset Selection becomes an important preprocessing part in the area of data mining. Hence, to reduce the dimensionality of the feature space, and to improve the performance, document preprocessing, feature selection and attribute reduction becomes an important parameter. To overcome the problem of document preprocessing, feature selection and attribute reduction, a theoretic framework based on hybrid Information gain-rough set (IG-RS) model is proposed. In this paper, firstly the document preprocessing is prepared; secondly an information gain is used to rank the importance of the feature. In the third stage a neighborhood rough set model is used to evaluate the lower and upper approximation value. In the fourth stage an attribute reduction algorithm based on rough set model is proposed. Experimental results show that the hybrid IG-RS model based method is more flexible to deal with documents.

Keywords: Document Preprocessing, Feature Selection, Information Gain, Roughset.

1 Introduction

Now a day the number of text document on the internet is increasing tremendously. To deal the large amount of data, data mining becomes an important technology. Text documents are growing rapidly due to the increasing amount of information available in electronic and digitized form [1][2], such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web. Document Preprocessing and feature selection is very useful tool in today's world where large amount of documents and information are stored and retrieved electronically. To solve this problem of dimensionality different technique such as document preprocessing, feature selection and attribute reduction approaches are used. Document preprocessing is a process that extracts a set of new terms [3][4] from the original document/ terms into some distinct key term set. Feature selection process a subset which selects from the original set based some criteria of feature importance. In a

S.C. Satapathy et al. (eds.), *Emerging ICT for Bridging the Future – Volume 2,*
Advances in Intelligent Systems and Computing 338, DOI: 10.1007/978-3-319-13731-5_59

wide range of text categorization many feature selection methods are used. [14] Proposed information gain is the most effective method compared to other five feature selection methods [16] such as IG, term strength, mutual information, X^2 statistic, document frequency.

Feature Selection, called as attribute reduction is a major problem in the field of data mining, pattern recognition and machine learning. Features may be relevant or irrelevant; it may have different discriminatory or predictive power. To apply measure to calculate uncertainty from fuzzy approximation spaces and [14][15] used it to reduce heterogeneous data. However, it becomes time consuming to generate a fuzzy equivalent relation. Therefore to be more effective a hybrid method information gain and Neighborhood Rough set model has been proposed.

2 Document Preprocessing

To Organize and browse thousands of documents smoothly, document preprocessing becomes a most important step, which affects on the result [5] where thousands of words are present in a document set; the aim of this is to reduce dimensionality for having the better accuracy for classification. Document preprocessing is divided into following stages:

1. Each sentences gets divided into terms
2. Stop words removal
3. Word Stemming
4. WordNet Senses
5. Global Unique words and frequent word set gets generated.

3 Feature Selection

Feature selection becomes an important task in machine learning, pattern recognition and data mining. It focuses on most important relevant features instead of irrelevant features [7] which makes more difficult in knowledge discovery process. Feature subset selection finds an optimal subset feature of a database based on some criteria, so that an efficient classifier with a highest accuracy can be generated. For text categorization [13][14] compared other five feature selection method which includes Information Gain (IG), X2 statistic document frequency, term strength, and mutual information. [15][16] Proposed that IG is the most effective method as compared to other feature selection method.

3.1 Feature Ranking with Information Gain

Information Gain is used as a significance measures based on entropy. For feature selection information gain is used which constitutes a filter approach. Information gain is an attribute selection measure and is based on entropy. Feature selection

depends on the IG value of the feature; it also determines which feature to be selected. In machine learning field information gain is the most popular feature selection method. The information gain of a given feature t_k with respect to the class c_i is the reduction in uncertainty about the value of c_i when we know the value of t_k Information gain of a feature tk toward a category c_i is labeled as follows

$$IG\ (\ t_k\ ,\ c_i\ ) = \sum_{c\ \epsilon\ \{c_i\ ,c_i\}} \sum_{t\epsilon\ \{t_k,\ t_k\}} P\ (\ t, c) \log \frac{p(t,c)}{p(t)p(c)}$$

Where $p(c)$ is the probability that category c occurs $p(t,c)$ is the probability that documents in the category c contains the word t. $p(t)$ is the probability that the term t occurs. The larger information gain of a feature owns the more important the feature is for categorization using WordNet.

3.2 Rough Set Theory

In this section we review several basic concepts in rough set theory and attribute reduction. Throughout this paper we suppose that the universe data used is denoted by information system $(IS) = < U, A >$, where U is a non empty and finite set of samples $\{x_1, x_2, x_3, .., x_n\}$ called an universe. A is a set of attributes to characterize the samples. $< U, A >$ is called as decision table. If $A = C \cup D$, where C is an condition attribute and D is an decision attribute. For Example given an arbitrary variable xi Є U and B $\subseteq$ C, the neighborhood $\delta_B\ (\ x_i\ )$ of x_i in feature space B is defined as

$\delta_B\ (\ x_i\ ) = \{x_j | x_j\ \in \cup,\ \Delta^B\ (x_i\ ,\ x_j) \leq \delta\}$, where Δ is a distance function. For $\forall_{x1,x2,x3}\ \in U$, it usually satisfies:

1. $\Delta\ (x_1, x_2\ ) \geq 0\ , \Delta\ (x_1, x_2\ ) = 0$ if and only if x1=x2;
2. $\Delta(x_1, x_2\ ) = \Delta\ (x_2, x_1\ )$;
3. $\Delta(\ x_1, x_3\ ) \leq \Delta\ (x_1, x_2\ ) + \Delta\ (x_2, x_3\ )$; The three different metric distance functions are most widely used in machine learning and pattern recognition. A general metric names minkowsky distance can also be defined.

If the set of objects and the neighborhood relation N over U is called as neighborhood approximation space. For any $\boldsymbol{X} \subseteq \boldsymbol{U}$, two objects called lower and upper approximation of x in $< \boldsymbol{U}, \boldsymbol{N} >$ are defined as

$\underline{\boldsymbol{N}}\boldsymbol{X} = \{\boldsymbol{x_i} \mid \boldsymbol{\delta}\ (\ \boldsymbol{x_i}\ ) \subseteq \boldsymbol{X}, \boldsymbol{x_i}\ \in\ \boldsymbol{U}\}$,

$\overline{\boldsymbol{N}}\boldsymbol{X} = \{\boldsymbol{x_i} \mid \boldsymbol{\delta}\ (\ \boldsymbol{x_i}\ ) \cap \boldsymbol{X}\ \neq\ \emptyset\ , \boldsymbol{x_i}\ \in \boldsymbol{U}\ \}$

Obviously $\underline{\boldsymbol{N}}\boldsymbol{X} \subseteq \boldsymbol{X}\ \subseteq \overline{\boldsymbol{N}}\boldsymbol{X}$. The boundary region of X in the approximation space is defined as $\boldsymbol{BNX} = \overline{\boldsymbol{N}}\boldsymbol{X} - \underline{\boldsymbol{N}}\boldsymbol{X}$. The size of the boundary effects on the degree of roughness of X in the approximation space$< \boldsymbol{U}, \boldsymbol{N} >$. The size of the boundary region depends on attribute X to hold U and threshold $\boldsymbol{\delta}$.

3.3 Decision System Based on Rough Set

An neighborhood information system also called a neighborhood decision system denoted by $NDT = <\ U, C\ U\ D, N >$, if there are two kinds of attribute: condition and decision, and there at least exists a condition attribute for the neighborhood relation.

Definition 1: Consider the neighborhood decision system $NDT = < U, C\ U\ D, N >$, $x_1, x_2, .., x_n$ are the objects with decisions 1 to N, $\delta_B(x_i)$ is the neighborhood information granule generated by attribute $B \subseteq C$, the lower and upper approximations of decision D with respect to attribute B defined as

$\underline{N_B}\ D = \cup_{i=1}^{N} \underline{N_B} X_i$, $\overline{N_B} D = \cup_{i=1}^{N} \overline{N_B} X_i$ Where $\underline{N_B}\ X = \{x_i | \delta_B (xi) \subseteq X, xi \in U\}$

$\overline{N_B}\ X = \{x_i | \delta_B (x_i) \cap X \neq \emptyset, x_i \in U\}$

The decision boundary region of D with respect to attribute B is defined as

$BN (D) = \overline{N_B}\ D - \underline{N_B}\ D$

Definition 2: Given any subset $X \subseteq U$ in neighborhood approximation space <U, A, N, we define variable precision Lower and upper approximation of X as

$\underline{N^k}\ X = \{x_i | I (\delta (x_i), X) \geq k, x_i \in U\}$, $\underline{N^k} X = \{x_i | I(\delta (x_i), X) \geq 1 - k, x_i \in U\}$, Where $1 \geq K \geq 0.5$.

Definition 3: Given the $NDT =< U, C \cup D, N >$, the distance function Δ and neighborhood size δ, the dependency degree of D to B is defined as $\gamma_B D = \frac{|POS_B (D)|}{|U|}$ Where $|\cdot|$ is the cardinality of a set. $\gamma_B(D)$ is the ability of B to approximate D. As $POS_B (D) \subseteq U$, we have $0 \leq \gamma_B(D) \leq 1$. we say D completely depends on B and the decision system is consistent in terms of Δ andδ. If $\gamma_B(D) = 1$; otherwise, it can be D depends on B in the degree of γ.

Definition 4: Given a neighborhood decision system $NDS =< U, C \cup D, N >, B \subseteq C, \forall a \in B$, it defines the significance of a in B as $Sig_1 (a, B, D) = \gamma_B(D) - \gamma_{B-a}(D)$ The significance of an attribute is based on three variables: a, B, and D. The above definition is for backward feature selection. In this redundancy features are eliminated from original feature one by one. The significance measure for forward selection is $Sig_2 (a, B, D) = \gamma_{BUa}(D) - \gamma_B(D)\ \forall a \in A - B$ As $0 \leq \gamma_B(D) \leq 1$ and $\forall a \in B : \gamma_B(D) \geq \gamma_{B-a}(D)$, we have $0 \leq Sig_1(a, B, D) \leq 1, 0 \leq Sig_2(a, B, D) \leq 1$. As attribute a is superfluous in B with respect to D if $Sig_1 (a, B, D) = 0$, otherwise, a is indispensible in B. With these proposed measure a forward greedy search algorithm for attribute reduction based on rough set theory has been proposed as follow

From the Algorithm 1 it shows that the positive region of decision becomes monotonous with the attribute. So to increase the speed a fast forward algorithm has been proposed. An algorithm named fast forward algorithm has been proposed and explained as follow.

Algorithm 1. Forward greedy search algorithm for attribute reduction based on rough set theory.

Input: 1. <U, C U D, F> 2. Delta 3. Size of the neighborhood

Output: reduct R

1. $\emptyset \rightarrow R$; //R contains all selected attributes
2. For each $a_i \in C - R$
3. Compute $\gamma_{R \cup a_i}(D) = \frac{\left|POS_{B \cup a_i}(D)\right|}{|U|}$
4. Compute $SIG(a_i, R, D) = \gamma_{Rua}(D) - \gamma_R(D)$
5. End. Consider the attribute a_k satisfying $SIG(a_k, R, D) = \max_i (SIG(a_i, R, D))$
6. If $SIG(a_k, R, D) > €$, where € is the positive number used for convergence.
7. $R \cup a_k \rightarrow R$ Goto step 2 Else Return R
8. End if

Algorithm 2. Fast forward attribute reduction based on rough set.

Input: <U,C U D> Delta, the size of the neighborhood

Output: reduct R

1. $\emptyset \rightarrow R, U \rightarrow S$; R is the selected attributes and S is the set of samples out of positive region.
2. While $S \neq \emptyset$
3. for each $a_i \in A - R$ Generate an decision table $DT_i = < U, R \cup a_i, D >$
4. $\emptyset \rightarrow POS_i$ for each $O_j \in S$ compute delta in NDT
5. if delta belongs to Xk $POS_i \cup O_j \rightarrow POS_i$
6. End if
7. End for
8. Find a_k If $POS_k \neq \emptyset$
9. $R \cup a_k \rightarrow R$ $S \rightarrow POS_k \rightarrow S$
10. Else
16. Exit While
17. End if
18. End While
19. return R
20 End

To consider discrete subsets an algorithm is modified as:

Algorithm 3. Fast forward discrete attribute reduction based on rough set.

Input: <U,C U D>, Delta, the size of the neighborhood
Output: reduct R
1. $\emptyset \rightarrow R, U \rightarrow S$; R is the selected attributes and S is the set of samples out of positive region.
2. While $S \neq \emptyset$
3. for each $a_i \in A - R$
4. Generate an decision table $DT_i = < S, R \cup a_i, D > \emptyset \rightarrow POS_i$
5. for each $O_j \in S$ compute delta in NDT
6. if delta belongs to Xk POS_i U $O_j \rightarrow POS_i$
7. End if
8. End for
9. Find a_k If $POS_k \neq \emptyset$
10. $R \cup a_k \rightarrow R$ $S \rightarrow POS_k \rightarrow S$
11. Else
12. Exit While
13. End if
14. End While
15. Return R
16. End

4 Experimental Analysis

The data sets used in the experiments is outlined in Table 1. All the experiments have been carried out on a personal computer with Windows 7, Inter(R) Core (TM) i7 CPU (2.66 GHz) and 4.00 GB memory. The software used is MATLAB R2010b. .

Table 1. The general description of Reuters 21578 dataset

Data set	Feature samples	Numerical attributes	Class
Reuters 21578	1328	24818	04

Table 2 shows the number of selected features where the information gain feature selection technique has applied.

Table 2. Feature reduced based on Information Gain

Data set	Feature samples	Numerical attributes	Information Gain	% of features reduced
Reuters 21578	1328	24818	9167	36.93

Table 3. shows the number of feature selected for Rough set by considering the delta interval from [0.001, 0.015,0.01]

Table 3. Number of Features Selected for Rough set

Data set	Algorithm	No. of Feature samples	Numerical attributes	RS		
				δ=0.001	δ =0.015	δ=0.01
Reuters 21578	Algorithm 1	1328	24818	497	483	477
	Algorithm 2	1328	24818	523	517	509
	Algorithm 3	1328	24818	538	524	518

Figure.1 presents the number of features selected for rough set approach.

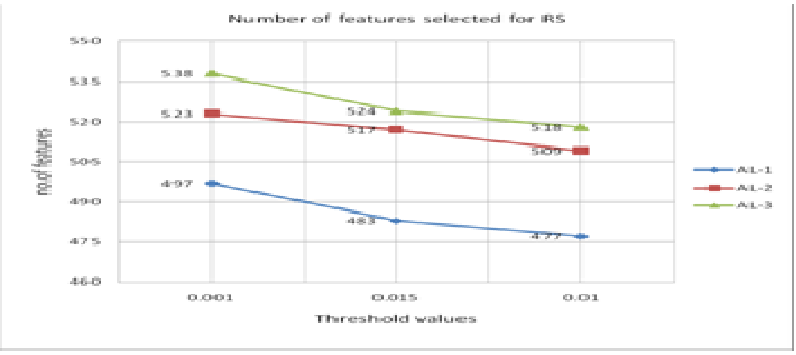

Fig. 1. Number of Features selected for Rough set

Table 4 shows the number of features selected using IG-RS technique for delta interval of [0.001, 0.5].

Table 4. No. of Features selected for IGRS

Data set	Algorithm	No. of Feature samples	Numerical attributes	IGRS		
				δ =0.001	δ =0.015	δ =0.01
Reuters 21578	Algorithm 1	1328	24818	115	110	98
	Algorithm 2	1328	24818	117	114	109
	Algorithm 3	1328	24818	121	117	111

Figure 2. Shows the number of features selected for IG-RS.

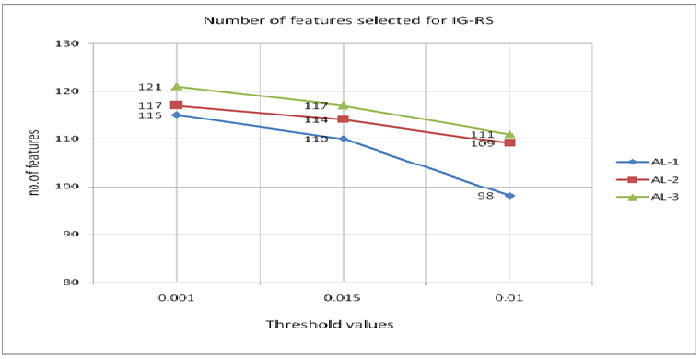

Fig. 2. Number of features selected for IGRS

5 Conclusion

In this paper, document preprocessing, feature selection, and attribute reduction approaches are used to reduce the high dimensionality of feature space composing the large number of terms. In this firstly the document preprocessing is performed where the stop words are removed, stemming is done, and global unique words are generated with the help of Wordnet. In the second stage feature selection approach called information gain is used to rank the importance of the features. Thirdly, an neighborhood rough set model is used to compute the lower and upper approximation value. Lastly attribute reduction algorithms are applied on rough set approach. In this a hybrid approach IG-RS shows the superior performance for attribute reduction feature selection.

References

[1] Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, KDD (2000)

[2] Fung, B., Wang, K., Ester, M.: Hierarchical document clustering using frequent item sets. In: Proc. of SIAM Int'l Conf. on Data Mining, SDM, pp. 59–70 (May 2003)

[3] Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proc. of Int'l Conf. on Knowledge Discovery and Data Mining, KDD 2002, pp. 436–442 (2002)

[4] Chen, C.L., Tseng, F.S.C., Liang, T.: An integration of fuzzy association rules and WordNet for document clustering. In: Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 147–159 (2009)

[5] Chen, C.-L., Tseng, F.S.C., Liang, T.: An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data & Knowledge Engineering 69, 1208–1226 (2010)

[6] Chen, C.-L., Tseng, F.S.C., Liang, T.: Mining fuzzy frequent itemsets for hierarchical document clustering. Information Processing and Management 46, 193–211 (2010)
[7] Yang, J., Liu, Y., Zhu, X., Liu, Z., Zhang, X.: A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing and Management (2012)
[8] Xu, Y., Wang, B., Li, J.-T., Jing, H.: An Extended Document Frequency Metric for Feature Selection in Text Categorization. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 71–82. Springer, Heidelberg (2008)
[9] Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning, pp. 412–420 (1997)
[10] Pawlak, Z., Skowron, A.: Rough Sets: Some Extensions. Information Sciences 177, 28–40 (2007)
[11] Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. IEEE Transactions of Knowledge and Data Engineering 16, 1457–1471 (2004)
[12] Jensen, R., Shen, Q.: Fuzzy-rough sets assisted attribute selection. IEEE Transactions on Fuzzy Systems 15(1), 73–89 (2007)
[13] Uğuz, H.: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems 24, 1024–1032 (2011)
[14] Hu, Q., Yu, D., Liu, J., Wu, C.: Neighborhood rough set based heterogeneous feature subset selection. Information Sciences 178, 3577–3594 (2008)
[15] Wang, H.: Nearest neighbors by neighborhood counting. IEEE Transactions on PAMI 28, 942–953 (2006)