Algorithm Study Based on Rough Entropy for Gene Analysis and Selection

Jiayang Wang   Zujian Wu

College of Information Science and Engineering, Central South University, Changsha 410083, P. R. China

Abstract

Gene expression data has been used to analyze and classify disease in resent years. Combining the attribute importance in Rough Sets Theory and entropy in Information Theory, this paper introduces the study of the gene analysis and selection method. A novel algorithm, called RMSME, is proposed to use the minimum uncertain information to reduct and generate the mostly related genes with the subclasses of disease. Finally, the experimental results show the effectiveness and practicability of this algorithm on the actual medical data.

Keywords: Rough set, Entropy, Bioinformatics, Gene

1. Introduction

After the successful progress of Human Genome Project, the latest research methods on genome are using the computer to analyse ,simulate and predict the new information,basing on the knowledge of biologic sequences and features.In bioinformatics, thousands of data of genes can be analysed on the high-throughput microarray platform.Therefore, many people are more interesting on how to choose the suitable theory to deal with these data sets from genes, for getting the genes which are most related with the diseases.

Rough Set Theory(RST) is a mathomatic theory to study the realistic information [1]-[2]. It focuses on knowledge reasoning by the whole set directly approximating to the incomplete and uncertain information.There are some studys on the konwledge reduction and features selection of mass data [3]-[5]. In this paper, a novel method using RST is offered to select genes from data set.

This paper is organized as follows.In section 2, there are some basic conceptions of RST;then a gene selection method named RMSME is described in details in section 3;after that, evaluation experiments and validity analysis are presented in section 4;finally,conclusions are listed in the last paragraph.

2. Theory

2.1. Significancy of attributes

The significancy of attributes can be described in the RST as follows. An information system, IS(or an approximation space),can be seen as a system :

$$IS = (U, A)$$

where U is the universe and A is the set of attributes. X is a subset($X = \{x_1, x_2, ..., x_m\}$), each attribute $x_i \subseteq A$. When x is added into the X, the ability of discerning objects of IS is enhanced and the enhanced degree is the significancy of attributes.If the degree is high,there is an opinion that x is important to X.

In an IS, C is condition attributes set and D is the decision attributes set.The significancy of subset $C' \subseteq C$ to the D is defined as [6]:

$$SIG_{CD}(C') = \gamma_C(D) - \gamma_{C-C'}(D)$$

where $\gamma_C(D) = |POS_C(D)| / |U|$, especially when $C' = \{a\}$, the significancy of attribute $a \in C$ to the D is written as :

$$SIG_{CD}(a) = \gamma_C(D) - \gamma_{C-\{a\}}(D)$$

2.2. Rough information entropy

Because of the uncertain information in the real life, there must be some quantitative definitions to present the uncertain degrees. The conception of Entropy is from the physics,presenting the mean information of knowledge. Entropy reflects the uncertain degree of information in knowledge set. Shannon used the conception of Information Entropy in information theoryand defined the basic measurement to the uncertain knowledge [7].There are some introductions to the Information Entropy and Condition Entropy [8]-[9].

U is the universe, $U = \{X_1, X_2, ..., X_n\}$, there is a probability distribution and information entropy of X is presented as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

where $p_i = P(X_i) = |X_i| / |U|$,and there are some propertys of $H(X)$:

(1) $0 \le H(X) \le \log n$

(2) if $X_i \subseteq U$ and $P(X_i) = 1$, then $H(X) = 0$

(3) when $p_1 = p_2 = ... = p_n = 1/n$, $H(X) = \log n$ is the maximum value.

If there is a $U = \{Y_1, Y_2, ..., Y_m\}$, $P(Y_j) = q_j$, $\sum_{j=1}^{m} q_j = 1$, then the condition entropy $H(Y|X)$ is defined as :

$$H(Y|X) = \sum_{i=1}^{n} P(X_i) H(Y|X_i)$$

where

$$H(Y|X_i) = \sum_{j=1}^{m} P(Y_j|X_i) \log P(Y_j|X_i)$$

$$P(Y_j|X_i) = |Y_j \cap X_i| / |X_i|.$$

There are also other defines about rough information entropy. Researchers described the conceptions from RST and Information Theory with different opinions.

According to the physical meaning of $H(P)$ which are mean information, mean uncertain and randomicity,in paper [10], the rough information entropy $E(P)$ is defined as：

$$E(P) = -\sum_{i=1}^{n} P(X_i) \log w_i$$

where w_i is the reciprocal of number of X_i set, namely $w_i = 1/card(X_i)$. This definition reflected the rough information of knowledge: when the rough information of knowledge is smaller and smaller,the offered information is much plenty .The mean uncertain and randomicity are less with small entropy value.

In paper [11], the conception of uncertain border-domain in RST is combined with the Entropy $H(X)$ in Information Theory.And there are definition and proofing of the relationship between rough information entropy H_B with $U_B(X)$:

$$H_B = U_B(X)$$

where

$$H_B = -\sum_{i=1}^{n} p(B_i) \log p(B_i)$$

$$U_B(X) = |R^{-}(X) - R_{-}(X)| / |U|.$$

This definition is based on the conception of lower and upper approximations in RST. Further more,the relation of RST and traditional probability statistics is studied. There is a conclusion that rough information entropy equals to traditional information entropy.Therefore, because of the reasonable way choosing the attribute point in a decision tree by simple lower and upper approximations, the study is simple than the traditional method of entropy computing which is basing on probability statistics.

There is also a new way to compute the entropy by RST on uncertain information [12].IF there is an information system $IS = (U, A)$, with $X \subseteq U$ and $P \subseteq A$, then the rough entropy of X to the knowledge P is defined as:

$$E_P(X) = \alpha_R(X) H(P)$$

where $\alpha_R(X)$ is an accuracy measure of the set X in R , $H(P)$ is the information entropy of knowledge P.

3. Gene analysis and selection

There are mainly statistics and information theory on the gene analysis and selection [13]-[14].Gene expression sequence data is the main data sources of the researchers and it can be presented by matrix format. The rows are for the genes number and the columns are for the samples in a matrix.

There are two types of gene features selection approaches:filters and wrappers [15]-[16].In filter type, characters selection is based on the relation of data and dividing of classes.Further more,filter type is an approach with more easy computing and high effectio, especially uncorrelated to that of the learning methods. In wrapper type methods,the characters selection is correlated with some learning methods.The utility of a character can be directly judged by the estimated accuracy of the learning method. Then, there is a subset of non-redundant characters with small numbers and high predicting rate.But there is much more computing for the better set of characters in wrapper type.

In paper [17], there is a RMIMR algorithm which was proposed by the researchers,basing on the conception of dependency of attributes in RST. The RMIMR algorithm is used to deal with the redundency data in trainnig period of classifier.Defining the relevance and interaction of gene,the paper analyzed the genes in information system which is componented by the genes expression data and disease smaples.After that, the reducted genes are offered to the training of the SVM classifier and $Naive-Byes$ classifier as training data set.There were successful experiments on the four different cancer data,using this algorithm and study method.

The main contribution of this paper is that we define significancy and rough entropy of genes basing on the rough set theory, and propose the method call RMSME(Rough Max SIG-Min Entropy), which is verified to be effective and useful by analysis and experiments.

3.1. Criterion of gene selection

The basic way to choose the genes is selecting the genes which are most related to the disease or having the most information on the gene expression sequence data.And we can sort the genes according to predefined standard.However,there are thousands of redundant genes in the medical database. Therefore,the aim of gene selection is to reduce theose genes that are irrelevant to the calssification of disease.This papre will use the significancy of attributes in RST and entropy in information theory to reduce the redundant genes ,estimate and get the relevant subset of genes.

Definition 1. Gene expression data contains n genes and a class variable D, the gene set is denoted by $Gene=(gene_1,gene_2,...,gene_n)$, and the class variable is denoted by $D=(D_1,D_2,...,D_m)$, U is the universe of the data, $|U|$ is the cardinality of U.The significancy of $gene_i$ can be written as:

$$S_{Gene}(gene_i)=(|POS_{Gene}(D)|-|POS_{Gene-\{gene_i\}}(D)|)/|U| \qquad (1)$$

where $|POS_{Gene}(D)|$ is the number of the positive region, $S_{Gene}(gene_i)$ is standing for the significancy of the ith gene to $Gene$ in the whole universe.

In the matrix of gene expression data,distinguishing the importance of gene can be computed by the $S_{Gene}(gene_i)$. In RST,the significancy of gene presents the relationships between the attributes and decision classification in information system.When the significancy of gene is much more than other genes, this gene has a much close relationship to the samples classification than other.And this principle is one of the important standards to the redundant attributes selection in the reduction.

Definition 2. Gene expression data contains n genes and and a class variable D , $Gene=(gene_1,gene_2,...,gene_n)$, $D=(D_1,D_2,...,D_m)$. $gene_i$ is the ith gene in $Gene$ and D_j is the jth classification in D. ($(i=1,2,...,n)$, $(j=1,2,...,m)$) . The condition entropy of gene$_i$ can be written as:

$$H(D|Gene)=-\sum_{j=1}^{m}P(D_j|\{gene_i\})\log P(D_j|\{gene_i\}) \qquad (2)$$

where $P(D_j|\{gene_i\})=|D_j\bigcup\{gene_i\}|/|\{gene_i\}|$ The $H(D|Gene)$ is standing for the gained uncertain information of class D from $gene_i$.

when $U/\{D\bigcup gene_i\}=\varnothing$,then $H(D|Gene)$ is the minimum condition entropy value of ith gene:

$$\min H(D|Gene)=0$$

when $U/\{D\bigcup gene_i\}=\{U\}$,then $H(D|Gene)$ is the maximum condition entropy value of ith gene:

$$\max H(D|Gene)=\frac{|\{U\}|}{|U|}\sum_{j=1}^{m}\frac{|\{U_m\}|}{|\{U\}|}\log\frac{|\{U_m\}|}{|\{U\}|}$$

From above maximum and minimum condition entropy values of $H(D|Gene)$, there is a standardization to the $H(D|Gene)$:

$$H^s(D|Gene)=1-\frac{H(D|Gene)}{\max H(D|Gene)} \qquad (3)$$

where $0\le H^s(D|Gene)\le 1$ and

(1) when $H^s(D|Gene)=0$,there is

$$gene_i\xrightarrow{H(D|Gene)=\max H(D|Gene)}D$$

D is completely depending on gene$_i$;

(2) when $H^s(D|Gene)=1$, there is

$$gene_i\xrightarrow{H(D|Gene)=\min H(D|Gene)}D$$

D is independing on gene$_i$.

According to the standardization of condition entropy,when $H^s(D|Gene)$ is close to the zero, there is less uncertain information.That is, much certain information is gained and the information dependance of $gene_i$ with D is much more stronger.

In paper [18], there is a discussions and proofment on the incompleteness of the two existing definitions of attribute significance in details. A modified definition of the attribute significance based on the weighed sum was proposed.

Basing on the above study results, this paper defines the parameter of gene analysis and selection, Rough Entropy(RE). This parameter is computed by the condition entropy of genes in systems,being judged and offering r elements in a subset of Gene.

Definition 3. In gene expression data system, there is a $Gene=(gene_1,gene_2,...,gene_n)$,.And a class variable $D=(D_1,D_2,...,D_m)$ .The gene rough information entropy is:

$$RE(gene_i)=(1-\alpha)S_{Gene}(gene_i)+\alpha H^s(D|Gene) \qquad (4)$$

the $H^s(D|Gene)$ is the main judgement factor and $S_{Gene}(gene_i)$ is the adjective one, therefore $0.9\le\alpha\le 1$. this definition is helping to choose the genes which contain small value of rough entropy. These genes make up a related subset of the diseases.Further more,they can also be took as training data to the classifier in the following steps. Therefore,there are r genes （$0 \le r \le n$） in $Gene$ which satisfy:

$$\max RE(gene_i) \qquad (5)$$

That means to find out r genes,minimum values of the entropy in $Gene$, which satisfy:

$$H(D|Gene) \qquad (6)$$

The genes which related with the samples classified are in the subset $S=(gene_1, gene_2,..., gene_r)$.

3.2. Algorithm of gene selection

According to the criteria described above, a method named RMSME (Rough Maximum Significancy-Minimum Entropy) is proposed, it uses a simple heuristic algorithm to resolve the find the aimed genes. The algorithm of RMSME method is described as follows.

Algorithm RMSME

Input：Gene expression data contains n genes and a class variable, $Gene=(gene_1, gene_2,..., gene_n)$ and $D=(D_1, D_2,..., D_m)$.

Output：Gene subset with r genes is denoted by $S=(gene_1, gene_2,..., gene_r)$.

```
Step1：$S \leftarrow \varnothing$ ；
Step2：for $i=1$to $n$ do
           Calculate $S_{Gene}(gene_i)$ according
             to Eqs.1；
       end
Step3：rank the
         $S_1=\{S_{Gene}(gene_1),S_{Gene}(gene_1),...,S_{Gene}(gene_i)\}$
         by decreasing order ；
Step4：choose genes from $S_1$from the top one
         to the last one and calculate
         $H^s(D|Gene)$according to Eqs.2 and
         Eqs.3 ；
Step5：rank the $n$ numbers of $H(D|Gene)$
         in $S_2$ by increasing order;
Step6：while $|S|\le r$ do
         for $i=1$ to $n$ do
             if   selected $S_{Gene}(gene_i)$
                  from $S_1$and
                  $H(D|Gene)$ from $S_2$
                  satisfy Eqs.6；
             then $S \leftarrow S+\{gene_i\}$；
                  $Gene \leftarrow Gene-\{gene_i\}$；
         $|S| \leftarrow |S|+1$
         end
       end
```

4. Experiments and analysis

4.1. Data set

For evaluating the validity of algorithm RMSME, the algorithm is carried out on the gene expression data set of $Leukemia$ in this paper (more details about the data set and data discretization can be refered to Http://www.broad.mit.edu/cgibin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43) .

The $Leukemia$ data set is one of the wildly used data set in gene study in paper [19]. The details of data set in this paper is showed Table 1.

Dataset	Gene	DataSubSet	Sample	Class	
Leukemia	7129			ALL	AML
		Train	38	27	11
		Test	34	20	14

Table 1: Leukemia data set

There are 7129 genes and 72 samples(train and test) in this data set. The samples belong to two subclass of $Leukemia$: ALL and AML.The discrete data set can be used to verify the algorithm.

In traditional information system, the columns are for condition and decision attributes ,and that the rows are for the samples. The number of genes(attributes) is much more than the one of samples in $Leukemia$ data set,therefore the columns and rows can be used to present the samples and attributes,respectively. The details of train data in $Leukemia$ data set is in Table 2.

	S1	S2	…	S38
Gene1	A	P	…	A

Gene2	P	M	…	P
…	…	…	…	…
Gene7128	A	A	…	M
Gene7129	A	P	…	A
Class	ALL	ALL	…	AML

Table 2: Information system of train data set

There are three status for the relationship between samples and genes. $A(Absent)$ 、 $P(Present)$ 、 $M(Marginal)$ are each standing for the meanings that genes are not expression in samples ,expression in samples and could not be judged.

4.2. Results and analysis

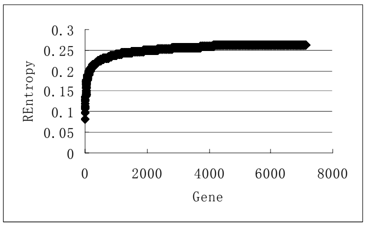
Based on the SQL language environment,the significancy of genes in S_1 are zero. S_1 can be described as:

$$S_1 = \{S_{Gene}(gene_1), S_{Gene}(gene_2), ..., S_{Gene}(gene_n)\}$$
$$= \{0, 0, ...0\}$$

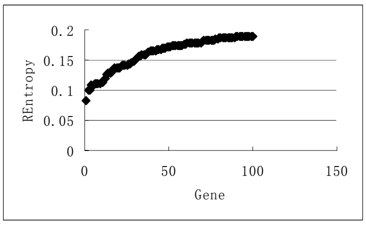
This result can be described as: when the number of genes is much more than the ones of samples.There is not obvious difference between the part genes as condition attributes with the whole genes as condition attributes.Therefore,the relationship between genes and samples could not be got just based on the attributes significancy.There must be more other conditions, such as entropy of genes, considered and used to analyze the relationship.

According to the definition of rough entropy,the rough entropys of each genes in S_2 can be computed. Fig.1 displays the computed results of the rough entropys.

(a)



(b)



(c)

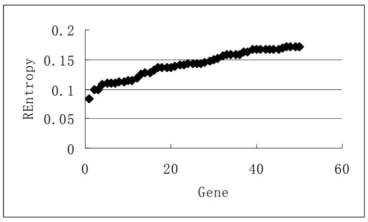


Fig. 1: Rough entropy of genes in S_2

From the (a) and (b) of Fig.1, there are small amount of genes which rough entropys are between 0.08 to 0.2, and most of 7129 genes with rough entropys are more than 0.2.The figures mean that few genes are relating to the classification of samples,and this conclution is consistent to the real results from medical data.

From the (c) of Fig.1, there are 20 genes in the genes set having the less uncertain information to the classification of samples.Thereby, the genes subset which contains genes related to the classification can be found by the foremer 20 genes in set. And the gene expression data of these 20 genes can be took as trainning data to train the classifier for samples classification. Finally,according to the classification accuracy, aimed genes can be selected. The *LibSVM* classifier is used to train the module and predict the samples classification [20]. The prediction results of samples in test data set of *Leukemia* are showed in Fig.2.

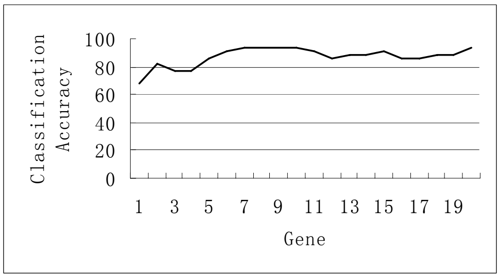


Fig. 2: Classification accuracy of test data set

The classification accuracy in Fig.2 shows that the 34 samples in test data set were classified by different number of genes as condition attributes and the figures ranges from 67.6471% to 94.1176%. the number of genes and corresponding classification accuracy to the samples are showed in Table 3.

Number of Gene	Classification Accuracy (%)	Number of Gene	Classification Accuracy (%)
1	67.6471	11	91.1765
2	82.3529	12	85.2941

3	76.4706	13	88.2353
4	76.4706	14	88.2353
5	85.2941	15	91.1765
6	91.1765	16	85.2941
7	94.1176	17	85.2941
8	94.1176	18	88.2353
9	94.1176	19	88.2353
10	94.1176	20	94.1176

Table 3: Genes and responding classification accuracy

There are 5 genes, in the fore 20 genes from above algorithm, consistent to the results of *Golub* et al.Table 4 gives the details of the 5 related genes.

Gene	Gene Accession Number	Rough Entropy	Gene Description
g1	M31211_s_at	0.108049573	MYL1 Myosin light chain (alkali)
g2	L47738_at	0.10977751	Inducible protein mRNA
g3	M84526_at	0.11145927	DF D component of complement (adipsin)
g4	Y12670_at	0.113572373	LEPR Leptin receptor
g5	U05259_rna1_at	0.13681769	MB-1 gene

Table 4 : Related genes and descriptions

From the above results,in Rough Set Theory,the rough condition information entropy takes the uncertain status of boundary field into account. When the uncertain information is less and less, the certain information of classification is more and more, by deducding the results from the knowledge conditions.

5. Conclusion

Following the Human Genome Project(HGP),the Post-genome research is focusing on the analysis of gene expression data by novel computing methods.And there are new theory frames on the bioinformatics area ,combining with the other theories.

Basing on the Rough Set Theory and the high dimension of gene expression data, the problems of gene selection are investigated in this paper.And a RMSME algorithm is proposed, based on the conceptions of rough set and information theory. According to the analysis and experiments, the results in this paper show that the most related genes are found out and much valued data is offered to the further researchs.

Acknowledgement

This work is supported by Hunan Natural Science Foundation (Grant No. 06JJ20075).

References

[1] Z. Pawlak, Rough sets. *International Journal of Computer and Information Sciences*, 11:341-356, 1982.

[2] Z. Pawlak, Rough sets. *Theoretical Aspects of Reasoning about Data*, Boston: Kluwer Academic Publishers, 1991.

[3] J.Y. Wang, S.Q. Chen, A. Luo, Study for dynamic reduct based on rough set. *MINI-MICRO SYSTEMS*, 27(11):2056-2060, 2006.

[4] M. Quafafou, M. Boussouf, Generalized rough sets based feature selection, *Intelligent Data Analysis*, 4:3-17, 2000.

[5] J.C. Han, X.H. Hu, T.Y. Lin, Feature subset selection based on relative dependency of attributes. *Rough Sets and Current Trends in Computing:4th International Conference, Uppsala, Sweden*, pp.176-185, 2004.

[6] M. He, B.Q. Feng, Z.F. Ma, X.H. Fu, Heuristic algorithm for reduction of attributes based on rough set theory. *MINI-MICRO SYSTEMS*, 26(3):356-359, 2005.

[7] C.E. Shannon, A mathematical theory of communication. *The Bell System Technical Journal*, 27:379-423, 1948.

[8] W.X. Zhang, W.Z. Wu, J.Y. Liang, D.Y. Li, Rough set theory and approach, *Beijing：Science Press*, 2001(in Chinese).

[9] J.Y. Liang, X.W. Meng, Application of information entropy in rough set theory. *Journal of Shanxi University(Nat.Sci.Ed.)*,25(3):281-284, 2002.

[10] Y.R. Li, B. Qiao, J.P. Jiang, The representation of the rough entropy of uncertainty in the rough set theory. *Computer Science*, 29(5):101-103, 2002.

[11] H. Zhu, Research of ID3 algorithm based on rough set. *Natural Science Journal of Xiangtan University*, 28(1):33-36, 2006.

[12] X.Y. Wang, N. Cai, J. Yang, X.J. Liu, A new method for measuring uncertainty in rough sets. *Journal of Shanghai Jiaotong University*, 40(7):1130-1134, 2006.

[13] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data. *IEEE Computer Society Bioinformatics Conference*, pp.523-529, 2003.

[14] D.V. Nguyen, D.M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*,18:1216-1226, 2002.

[15] R. Kohavi, G.H. John, Wrapper for feature subset selection. *Artificial Intelligence*, issue 1-2, pp.273-324, 1997.

[16] P. Langley, Selection of relevant features in machine learning. *in AAAI Fall Symposium on Relevance*, 1994.

[17] D.F. Li, W. Zhang, Gene selection using rough set theory, *RSKT 2006, LNAI 4062*, pp.778-785.

[18] F. Shi, Z.L. Lou, Y.Q. Zhang, A modified heuristic algorithm of attribute reduction in rough set. *Journal of Shanghai Jiaotong University*, 36(4):478-481, 2002.

[19] T.R. Golub, DK Slonim, Molecular classification of cancer:class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

[20] http://www.csie.ntu.edu.tw/~cjlin/libsvm.