International Journal of Computing Algorithm
Volume: 01 Issue: 02 December 2012 Page No.38-41
ISSN: 2278-2397

Gene Selection Based on Rough Set

Applications of Rough Set on Computational Biology

K.Anitha
Department Of Mathematics, S.A.Engineering College, Chennai, India
Email: subramanianitha@yahoo.com

Abstract-Gene selection is a main procedure of discriminate analysis of microarray data which is the process of selecting most informative genes from the whole gene data base. This paper approach a method for selecting informative genes by using Rough Set Theory. Rough Set Theory is a effective mathematical tool for selecting informative genes. This paper describes basics of Rough Set Theory and Rough Set attribute reduction by Quick –Reduct based Genetic Algorithm.

Keywords-Rough Sets, Rough Set Attribute Reduction, Genetic Algorithm.

I. INTRODUCTION

The advent of microarray technology has meant that transcriptional responses to changes in cellular state can now be quantified for thousands of genes in a single experiment. Microarrays thus offer a window into transcriptional mechanisms underlying major events in health and disease.

Rough set theory, proposed in is a good mathematical ool for data representation and reduction. Its methodology is concerned with the classification and analysis of missing attribute values, uncertain or incomplete information systems and knowledge, and it is considered one of the first non-statistical approaches in data analysis . Any subset defined by its upper and lower approximation is called "Rough Set". The ideas of Rough Set proposed by Pawlak in 1980 and he is known to be 'Father of Rough Set Theory'

II. DIFFERENCES AMONG FUZZY , ROUGH AND CLASSICAL SET THEORY

In classical set theory a set is uniquely determined by its elements. In other words, it means that every element must be uniquely classified as belonging to the set or not. Lotfi Zadeh proposed completely new, elegant approach to vagueness called *FUZZYSET THEORY*. In his approach an element can belong to a set in a degree k ($0 \le k \le 1$), in contrast to classical set theory where an element must definitely belong or not to a set.

For example in classical set theory one can be definitely ill or healthy, whereas in fuzzy set theory we can say that someone is ill (or healthy) in 70 percent (i.e. in the degree 0.7). Rough set theory is still another approach to vagueness. Similarly to fuzzy set theory it is not an alternative to classical set theory but it is embedded in it. Rough set theory is still another approach to vagueness which is expressed by its boundary region not by its partial membership.

III. TERMINOLOGIES ON ROUGHSET THEORY

A. In discernibility Relation:

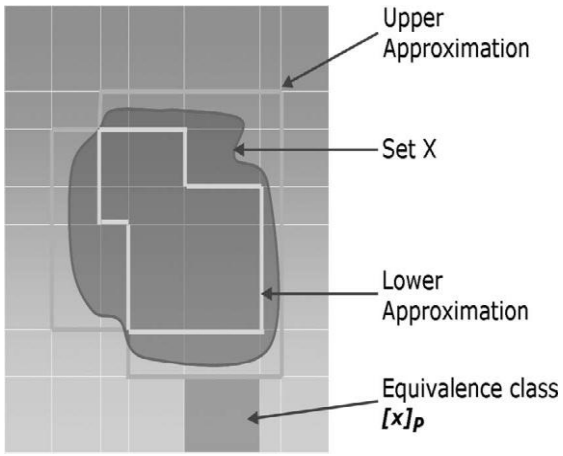
With any $P \subseteq A$, there is an associated Equivalence Relation $IND(P) = \{ (x , y) \in U \times U / \forall a \in P, a(x) = a(y)\}$.

B. Lower and Upper Approximation:

Let $X \subseteq U$ can by approximated using only the information contained within P, by constructing P-Lower and P-Upper approximations of a classical crisp set X are given by

$Px (Low) = \{ x / [x]p \subseteq X \}$
$Px (Upp) = \{ x / [x]p \cap X \neq \Phi \}$

The following figure shows diagrammatic representation of Rough set.



Using the features in the set P through equivalence Class we can construct Upper and Lower Approximations.

C. Feature Dependency and Significance

An important issue in data analysis is discovering dependencies between attributes. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \Rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P. If there exists a functional dependency between values of Q and P, then Q depends totally on P. In rough set theory, dependency is defined in the following way
For $P, Q \subset A$, it is said that Q depends on P in a degree k ($0 \le k \le 1$), denoted $P \Rightarrow k\ Q$, if
$k = \gamma P\ (Q) = |POS_P\ (Q)| / |U|$ ---------- Form(I) where $|S|$ stands for the cardinality of set S.

International Journal of Computing Algorithm
Volume: 01 Issue: 02 December 2012 Page No.38-41
ISSN: 2278-2397

D. Reducts

For many application problems, it is often necessary to maintain a concise form of the information system. One way to implement this is to search for a minimal representation of the original dataset. For this, the concept of a *reduct* is introduced and defined as a minimal subset *R* of the initial attribute set C such that for a given set of attributes *D*, $\gamma R(D) = \gamma C(D)$. Reducts are subsets of the attribute set A, which provide the same information as the original data set. The reducts were used as initial group centroids, which were then grouped together to form clusters.

Given an Information System A = (U,A)
we say that B ⊆ A of attribute is a Reduct of A
when
(i) INDB = IND A
(ii) B is a minimal set of attribute with property (i)

IV. DISCERNIBILITY MATRIX

Many applications of rough sets make use of discernibility matrices for finding rules or reducts. A discernibility Matrix of a decision table is a Symmetric matrix which entries defined by C ij = { a ∈ C / a (x i) ≠ a (x j) } Each Cij contains those attributes that differ between objects i & j.

A. Rough Set Attribute Reduction (RSAR)

Rough set attribute reduction (RSAR) provides a filter-based tool by which knowledge may be extracted from a domain in a concise way; retaining the information content while reducing the amount of knowledge involved. The main advantage that rough set analysis is that it requires no additional parameters to operate other than the supplied data . In RSAR a subset with minimum Cardinality is searched for the Original data set. Using the example dataset , in Table 1 the dependencies for all possible subsets of C can be calculated by Form(I). Table 1 consists of 8 objects with 4 conditional features a,b,c,d and one decision feature e. Given system is consistent if each set of object has same attribute value and whose corresponding decision features are same.

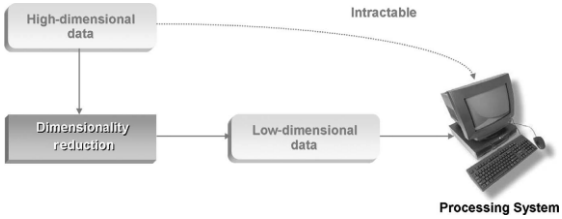
The possible selected subsets are given below:

- $\gamma\{a,b,c,d\}(\{e\}) = 8/8$ $\gamma\{b,c\}(\{e\}) = 3/8$
- $\gamma\{a,b,c\}(\{e\}) = 4/8$ $\gamma\{b,d\}(\{e\}) = 8/8$
- $\gamma\{a,b,d\}(\{e\}) = 8/8$ $\gamma\{c,d\}(\{e\}) = 8/8$
- $\gamma\{a,c,d\}(\{e\}) = 8/8$ $\gamma\{a\}(\{e\}) = 0/8$
- $\gamma\{b,c,d\}(\{e\}) = 8/8$ $\gamma\{b\}(\{e\}) = 1/8$
- $\gamma\{a,b\}(\{e\}) = 4/8$ $\gamma\{c\}(\{e\}) = 0/8$
- $\gamma\{a,c\}(\{e\}) = 4/8$ $\gamma\{d\}(\{e\}) = 2/8$
- $\gamma\{a,d\}(\{e\}) = 3/8$
- The given dataset is consistent, since

$\gamma\{a,b,c,d\}(\{e\}) = 1$. *The minimal* reduct set for this example is *Rmin = {{b, d}, {c, d}}* If {b,d} is selected the resultant reduced set is given in Table 2

B. Dimentionality Reduction

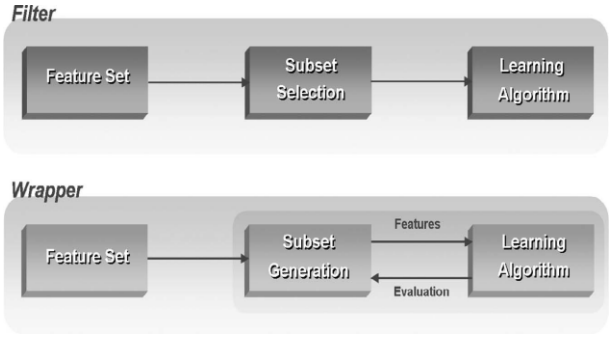
There are many factors that motivate the inclusion of a dimensionality reduction step in a variety of problem-solving systems . Many application problems process data in the form of a collection of real-valued vectors . If these vectors exhibit a high dimensionality, then processing becomes infeasible. For this reason it is often useful, and sometimes necessary, to reduce the data dimensionality to a more manageable size with as little information loss as possible and the process is described as follows



There are two main approaches in Dimentionality Reduction . They are

(i) Filter Method
(ii) Wrapper Method.

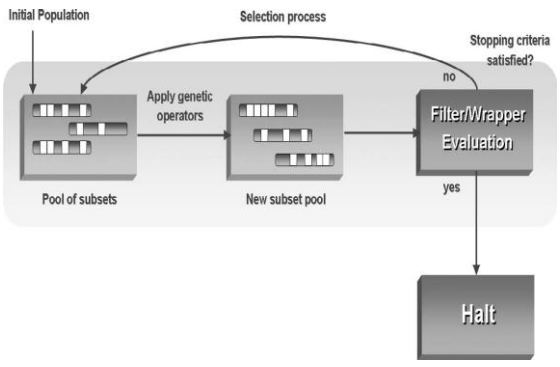
The first feature selection algorithms were based on Filter Approach. In this approach each feature is given a relevance weighting that reflects its ability to discern between decision class values. In Wrapper Method we will get intermediate solutions while working toward better ones that result in a lower classification error. This algorithm requires two threshold values to be supplied. The following diagram represents Filter & Wrapper Method Algorithms.



C. Gene Selection With Speciated Genetic Algorithm. (GA)

Genetic algorithms (GA) are generally effective for rapid search of large, nonlinear, and poorly understood spaces. Unlike classical feature selection strategies where one solution is optimized, a population of solutions can be modified at the same time. This can result in several optimal feature subsets as output.

D. Feature Selection With Genetic Algorithm



.

International Journal of Computing Algorithm
Volume: 01 Issue: 02 December 2012 Page No.38-41
ISSN: 2278-2397

E. Quick Reduct Algorithm

It is a process of finding all possible subsets of a given set. It starts off with empty set adding one at a time which gives the maximum attribute value in the rough dependency matrix.
The following algorithm is the sample Quick reduct based on wrapper method.
Quickreduct(C,D)

- **Input: C, the set of all conditional features; D, the set of decision features**

Output: *R, the feature subset*
(1) $R \leftarrow \{\}$
(2) **while** $\gamma_R(D) = \gamma_C(D)$
(3) $T \leftarrow R$
(4) **for each** $x \in (C-R)$
(5) **if** $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) **return** ***R***

GA is often used to select informative genes working together with classifiers to consider the mutual dependency among them .However, conventional wrapper methods using the GA are By using genetic algorithm we can construct efficient system for classification or decision making. The main idea of genetic algorithm based on Darwinian Principle of 'Survival of Fittest'. By using Rough Set based Genetic Algorithm we can calculate length of the found reduct and fittest function. It is used to select the most informative genes working together with classifier to consider the mutual dependencies. Gene Expression Data is obtained by extraction of quantitative information from the images/patterns resulting from the readout or fluorescent or radioactive hybridization in an Micro Array chip. The following Genetic Algorithm is based on Quick Reduct Algorithm.
The classical genetic algorithm we are given a state space S and f:S → R such that

$f(X0) = \text{Max}\{ f(x) / x \in S \}$

In GA a Chromosome is an n-element permutation γ, represented by sequence of numbers $\gamma(1), \gamma(2), \ldots \gamma(n)$.

Finding reduct based on permutation using genetic algorithm

INPUT.

(i) Decision table A = [U, {a1,a2,…an} U {d}]
(ii) Permutation γ generated by genetic algorithm.
OUTPUT : A Reduct R generated based on permutation γ

Method:
R = {a1,a2,…an}
(b1,b2,…bn) = γ (a1,a2,…an)
for i = 1 to n do
begin
R = R – b I
if not Reduct (R, A) then R = R U b end.

V. CONCLUSION

The approach gene selection based on Rough Set has a better performance than classical sets, for it avoiding the loss of information . In this paper we reduce a sample data set by using quick reduct algorithm .In same way we describe genetic algorithm for selecting informative genes. The result of algorithm will always be a reduct.

Table 1: Sample Data Set

X	A	B	C	D	E
0	S	R	T	T	R
1	R	S	S	S	T
2	T	R	R	S	S
3	S	S	R	T	T
4	S	R	T	R	S
5	T	T	R	S	S
6	T	S	S	S	T
7	R	S	S	R	S

Table 2: Sample Data Set

X	B	D	E
0	R	T	R
1	S	S	T
2	R	S	S
3	S	T	T
4	R	R	S
5	T	S	S
6	S	S	T
7	S	R	S

REFERENCES

[1] C. G. G. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for ForensicScientists*, 2nd ed. New York: Wiley. 2004.
[2] C. G. G. Aitken, G. Zadora, and D. Lucy. A two-level model for evidence evaluation. *J. Forensic Sci.* 52: 412–419. 2007. Forthcoming.
[3] C. G. G. Aitken, Q. Shen, R. Jensen, and B. Hayes. The evaluation of evidence for exponentially distributed data. *Comput. Stat. Data Anal.* 12(12): 5682–5693. 2007.
[4] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In *9th National Conference on Artificial Intelligence.* Cambridge: MIT Press, pp. 547–552. 1991.
[5] J. J. Alpigini, J. F. Peters, J. Skowronek, and N. Zhong, eds. *Rough Sets and Current Trends in Computing. Proceedings. 3rd International Conference*, Malvern, PA, October 14–16, 2002. Lecture Notes in Computer Science 2475. Berlin: Springer.
[6] 2002.
[7] K. K. Ang and C. Quek. Stock trading using RSPOP: a novel rough set-based neuro-fuzzy approach. *IEEE Trans. Neural Net.* 17(5): 1301–1315. 2006.
[8] C. Apt´e, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Sys.* 12(3): 233–251. 1994.
[9] S. Asharaf and M. N. Murty. An adaptive rough fuzzy single pass algorithm for clustering large data sets. *Pattern Recog.* 36(12): 3015–3018. 2004.
[10] S. Asharaf, S. K. Shevade, and N. M. Murty. Rough support vector clustering. *Pattern Recog.* 38(10): 1779–1783. 2005.
[11] J. Atkinson-Abutridy, C. Mellish, and S. Aitken. Combining information extraction with genetic algorithms for text mining. IEEE Intelligent Systems 19(3): 22–30.2004.
[12] *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*, by Richard Jensen and Qiang Shen Copyright © 2008 Institute of Electrical and Electronics Engineers

International Journal of Computing Algorithm
Volume: 01 Issue: 02 December 2012 Page No.38-41
ISSN: 2278-2397

[13] [11]. G. Attardi, A. Gull´ı, and F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis. In *Proceedings of 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105–119. 1999.
[14] W. H. Au and K. C. C. Chan. An effective algorithm for discovering fuzzy rules in relational databases. In *Proceedings of the 7th IEEE International Conference on Fuzzy Systems*. Piscataway, NJ: IEEE Press, pp. 1314–1319. 1998.
[15] T. Baeck. *Evolutionary Algorithms in Theory and Practice.* Oxford: Oxford University Press. 1996.
[16] A. A. Bakar, M. N. Sulaiman, M. Othman, and M. H. Selamat. Propositional satisfiability algorithm to find minimal reducts for data mining. *Int. J. Comput. Math.* 79(4): 379–389. 2002.
[17] 15. M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford. The isomap algorithm and topological stability. *Science* 295(5552): 7. 2002.
[18] [16]. J. F. Baldwin, J. Lawry, and T. P. Martin. A mass assignment based ID3 algorithm for decision tree induction. *Int. J. Intell. Sys.* 12(7): 523–552. 1997.
[19] [17]. J. K. Baltzersen. An attempt to predict stock market data: a rough sets approach.
[20] Diploma thesis. Knowledge Systems Group, Department of Computer Systems and Telematics, Norwegian Institute of Technology, University of Trondheim. 1996.
[21] [18]. P. Baranyi, T. D. Gedeon, and L. T. K´oczy. A general interpolation technique in fuzzy rule bases with arbitrary membership functions. In *Proceedings of IEEE*
[22] *International Conference on Systems, Man, and Cybernetics*. Piscataway, NJ: IEEE Press, pp. 510–515. 1996
[23] [19]. P. Baranyi, D. Tikk, Y. Yam, and L. T. K´oczy. A new method for avoiding abnormal conclusion for α-cut based rule interpolation. In *Proceedings of FUZZ-IEEE'99* ,Seoul, Korea. Piscataway, NJ: IEEE Press, pp. 383–388. 1999.