IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Improving Cancer Classification Accuracy Mistreatment Principle Element Analysis Methodology

R. Rosy Angel[1] and P. Subashree Kasithangam[2]

[1]**Assistant Professor, Department of Computer Science and Engineering, Holycross Engineering College, Tuticorin**

[2]**Assistant Professor, Department of Computer Science and Engineering, Holycross Engineering College, Tuticorin**

Abstract

Cancer classification enables the definition of therapeutic groups, for which therapeutic protocols can be elaborated, taking into account all treatment possibilities. Most classifications are based on clinical data. Most of the tumors have similar appearance so histological analysis tends to be unreliable. The advances in microarray technology make individualized treatment possible and when the classification of cancer is done based on minimal rough fringe. The result will be more accurate and reliable. This method dynamically evaluates all available genes and sifts the gene with small implicit regions of the dimension of the implicit hypercuboid. When an unseen object falls in the class hypercuboid then the object can be easily predicted. By vigorously constructing implicit hypercuboid, the approach selects the potential functional genes for suggesting classifiers. The experimental results on other classification such as that the proposed method can be faithful to the enhancement of the method for manipulating noises in data by employing variable precision rough sets. Hence Principle Component Analysis(PCA) is a feasible method for classifying cancer tissues and which uses proposed method for removing more uncertain region and reduces the number of implicit region.

Keywords:*Histological Analysis, Rough fringe, Implicit Region, Explicit Region*

1. Introduction

Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, hospitals and scientific discovery.

In bio – medical field, classification enables the definition of therapeutic groups, for which therapeutic protocols can be elaborated, taking into account all treatment possibilities. It is essential for physicians to establish the classification of a tumor before any treatment can be administered to the patient in order to avoid proposing unnecessary treatment (for instance mutilating surgery when the patient unfortunately has metastates) and to propose the most appropriate treatment (for instance: general treatment when localized treatment might be more appropriate). In this paper, a new ensemble approach is used for classifying cancer types based upon the rough sets theory.

Some of the already existing methods in the classification of cancer include Prediction Analysis for Microarrays (PAM) for analyzing gene expression data. It is a statistical method which induces classification rules based on the nearest shrunken centroids of different tumor types. In the training stage, the method performs cross validation to learn optimal amount of shrinkage, and hence, to sift genes for classification. Also simple methods, TSP and k – TSP are used for inducing decision rules based on the top or k-top scoring pairs of genes. The approaches are based on the concept of relative expression reversals of gene pair.

With the rapid progress of microarray technology many efforts have been being devoted to manipulating gene expression data. Distinguishing different tumors type based on microarray data involved in defining unrecognized tumors subtypes and classifying particular tumor samples into already defined samples. Various machine learning methods can be exploited for manipulating microarray data, which aims to cure the patient with an effective and safe drug. So, here the classification of cancer is done by using Rough sets theory. It is a mathematical tool to deal with vagueness and uncertainty. It was widely studied in

many fields such as machine learning, data mining, and pattern recognition.

2. Rough Sets Theory

Rough set theory was introduced by Polish Mathematician Pawlak in 1980s. It is regarded as a new mathematical tool to deal with vagueness and uncertainty. The theory has found in many domains, such as decision support, engineering, environment, banking, medicine and others. Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary – not competing discipline, in its own rights. Based on the rough sets theory, define explicit and implicit regions and introduced a new method for inducing decision trees in light of the principle of minimal rough fringe.

The set of all indiscernible (similar) objects is called an elementary set, and forms a basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as a crisp (precise) set – otherwise the set is rough (impre-cise,vague). Each rough set has boundary – line cases, i.e., objects which cannot be with certainly classified, by employing the available knowledge, as members of the set or its complement.

Rough set based data analysis from a data table called a decision table, columns which are labeled by attributes, rows – by objects of interest and entries of the tables are attribute values. Attributes of the decision table are divided into two disjoint groups called condition and decision attributes, respectively. Each row of a decision table induces a decision rule, which specifies decision (action, results, outcome, etc.) if some conditions are satisfied. If a decision rule uniquely determines decision in terms of conditions – the decision rule is certain. Otherwise the decision rule is uncertain. Decision rules are closely connected with approximations. Roughly speaking, certain decision rules describe lower approximation of decision in terms of conditions, whereas uncertain decision rules to the boundary region of decisions.

Classification of tumors is essential for successful treatment of cancer. By allowing the monitoring of expression levels for thousands of gene simultaneously, such techniques may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel approach to cancer classification.

Gene expression data are characterized by dimensionalities. It can be represented by a n × m matrix after preprocessing. Each row represents the expression levels of a gene across different biological samples; each column represents the gene expression levels of a genome under a sample, Usually n >> m, i.e., the number of variables(genes) is much greater than the number of biological samples. The number of sample is < 200 and the number of gene > 5000 generally. These data are not noise – free because their raw data contain a large amount of systematic noise and pre – processing algorithms cannot remove them completely. Although there are large amount of variables in these data, only a small set of variables have meaningful contributions to data variations.

For understandings the rationale of rough hypercuboid approach, review the definitions in the rough set theory [12], [13].

Given a knowledge representation system:

$$S = (U, Q, V, \rho) \quad (1)$$

where U is a certain set of objects called the universe, and Q denotes the set of attributes. It is usually divided into twosubsets, C and D, which denote the set of conditionattributes and the set of decision attributes, respectively.

ρ: U × Q ⟶ V is an information function, where V = Uaεq Va in which Va is the domain of attribute aεq.

For any subset G of C or D, an equivalence relation e G on U can be defined such that a partition of U induced by it can be obtained. Denote the partition as:

$$\underline{G}(X) = \bigcup_{Gi \subseteq X} Gi \quad (2)$$

The upper approximation of X on G˜ is

$$G^{-}(X) = \bigcup_{Gi \cap x \neq \phi} Gi \quad (3)$$

The negative region of X on G˜ is

$$NEG_G(X) = U - G(X) \quad (4)$$

The boundary region of X on G˜ is

$$BND_G(X) = G^{-}(X) - \underline{G}(X) \quad (5)$$

To avoid some misunderstandings of the definitions, the explicit and implicit regions are

defined. For a sample, if it falls within the implicit region, we cannot clearly assign a class label for it. This is called uncertainty. In real applications, one always tries to eliminate uncertainty as much as possible. If more uncertainty can be removed by choosing a gene or some genes, these genes will be chosen for classification.

3. Rough Hypercuboid Approach

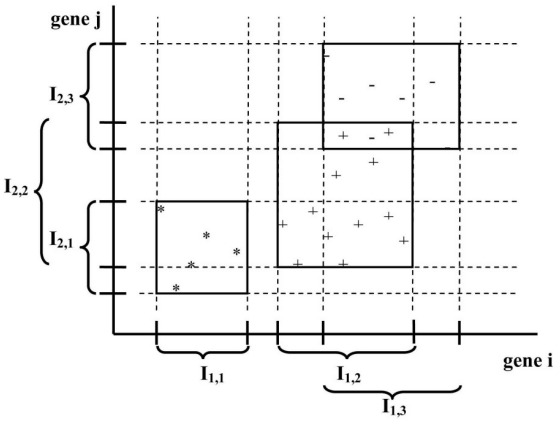
The concept of hypercuboid has been exploited in the machine learning society for many years. In this case description space consists of a setoff relation trees, an extension to their analysis in which the description space is constructed as a geometric space. Under this construal the behavior of both the Focusing algorithm and the classification algorithm is analyzed in terms of the construction of hypercuboid. This analysis leads to a number of observations: (i) that a distinction can be made between a strong and a weak version of the disjunctive – concept problem (ii) that certain solutions to the distinctive – concept problem can be shown to exploit what are, in effect, distance functions over the description space and (iii) that the classification algorithm is only capable of learning a subset of the possible disjunctive concepts in any given domain.

Generally, an n-dimension hypercuboid or hyper rectangle is defined in the n-dimension Euclidean space, where the space is defined by the n variables measured for each sample. The concept of hypercuboid has been exploited in the machine learning society. In the nearest hyper rectangle learning method, an unseen object can be predicted to the nearest hyper rectangle. All the nearest neighbor and nearest hyper rectangle algorithms share the common idea with the pervasively used.

For each gene gi, the values of its expression level of the m profiles are first rearranged into l intervals Ii;1; Ii;2; . . . ; Ii;l according to the cancer types of the profiles. Interval Ii;h is the value range of gene gi with respect to class Clh. It is spanned by the profiles with the same class label Clh. That is, the gene expression value of each profile with class label Clh falls within interval Ii;h. This can be viewed as a supervised discretization process, which utilizes class information.However, such a discretization process does not necessarilyresult in a compatible discretization , for every twointervals may intersect with each other. These intersectionswill form the so-called implicit hypercuboids. Tthe construction of rough hypercuboid classifiers is carried out by finding the implicit hypercuboids thatencompass the smallest number of misclassified profiles.

RHC Algorithm
Input: Training gene expression information system $S=(U, C \bigcup D, V, \rho)$.
The number of samples in U is m, the number of genes in C is n.
β is the error rate. The number of classes is l. Each class has been Labeled as $1,2, \ldots, l$. k is the dimension of rough hypercuboids.
Output: Classifer *RHC*

1. SET $Imp_A=U$, $Imp=\phi$, $k=0$;
2. FOR each gene g_i
   The value of gene g_i is rearranged into interval $I_{i,1}, I_{i,2}, \ldots, I_{i,l}$;
   END FOR
3. FOR each none selected gene g_i
   $Count(i)=0$;
   FOR each profile x in Imp_A
   FOR $h=1$ TO $l-1$
   FOR $f=2$ TO l
   IF $I_{i,h} \bigcap I_{i,f} \neq \phi$ AND $\rho(x, g_i) \in I_{i,h} \bigcap I_{i,f}$
   IF $k=0$ OR $HY_h^k \bigcap HY_f^k \neq \phi$
   $Count(i)=Count(i)+1$;
   END IF
   END IF
   END FOR
   END FOR
   END FOR
   END FOR
4. $j=\arg\min_i(Count(i))$, $Imp=Imp \bigcup \{(I_{j,1}, I_{j,2}, \ldots, I_{j,l})\}$;
5. $k=k+1$, $Imp_A=Imp_A-\{x | x \in Imp_A \wedge \forall(h,f), x \notin HY_h^k \bigcap HY_f^k\}$;
6. IF $Count(j) < \beta$, GOTO 7; ELSE GOTO 3;
7. RETURN *RHC*



From the figure, two – dimension hypercuboids the horizontal axis is gene i and the vertical axis is gene j. They are encoded as "1, 2," which indicate the first and second dimensions of the constructed hypercuboid. The three rectangles (solid line) are the three class hypercuboids with respect to class "_," "þ," and "_." They are simply encoded as "1," "2," and "3." Consequently, the rectangle I1;1 _ I2;1 at the left side is the class hypercuboid of class "_," denoted as HY 2 1 . The rectangle I1;2 _ I2;2 is the class hypercuboid of class "þ," denoted as HY 2 2. The rectangle I1;3 _ I2;3 is the class hypercuboid of class "_," denoted as HY 2 3 . A certain symbol "_" within the rectangle I1;1 _ I2;1 indicates a profile with class label "_." The values of gene i and j of this profile fall within I1;2 _ I2;2 and I1;3 _ I2;3, intersect with each other. In the intersection, there are two profiles with class label "þ" and one with "_." The intersection of two rectangles also forms a rectangle. It is referred to as

an implicit hypercuboid. The three profiles in the implicit hypercuboid form the implicit region.

It may be slightly strange to find that some profiles may simultaneously belong to different class hypercuboids. In fact, it is easy to understand that for a certain cancer type, a certain gene of various profiles always takes the gene expression values within a range. This range can be regarded as the value interval with respect to this cancer type. For a certain gene, different corresponds to different value ranges. If the ranges or intervals do not intersect with each other, this gene can be certainly used to distinguish different cancer types.

4. Principle Component Analysis (PCA)

When measuring only two variables it is easy to plot this data and to visually assess the correlation between these two factors. However, in a typical microarray experiment, the expression of thousands of genes is measured across many conditions such as treatments or time points. Therefore, it becomes impossible to make a visual inspection of the relationship between genes or conditions in such a multidimensional matrix. One way to make sense of this data is to reduce its dimensionally. Different cancers have different molecular patterns and the molecular patterns of a normal cell will be different from those of a cancer cell.

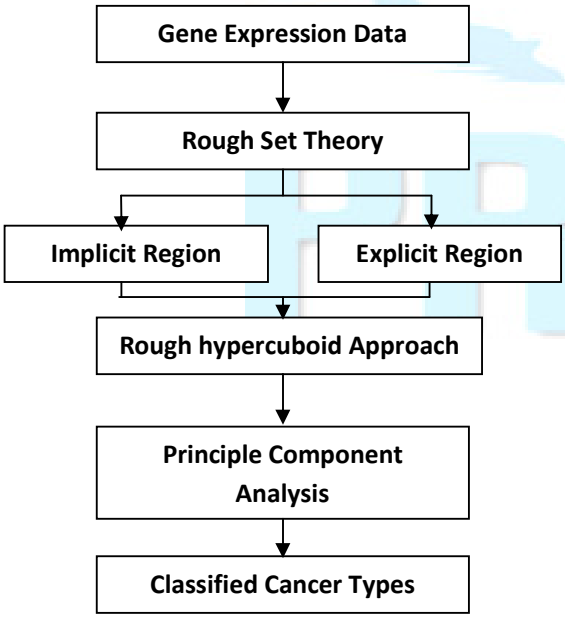


Fig 1: Flow Diagram For Proposed method

Objectives of principal component analysis are

- To discover or to reduce the dimensionality of the data set.
- To identify new meaningful underlying variables.

The mathematical technique used in PCA is called Eigen analysis. To solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalues has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalues determines the direction of the second principle component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

5. Conclusion

From the gene expression data set, the genes are classified if the constructed implicit region or implicit hypercuboid cluster contains the smallest number of profiles compared with that constructed by other genes. By dynamically constructing implicit hypercuboids, the approach selects potential functional genes for inducing classifiers. PCA method is used for removing uncertain region and used to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset. The results suggest that the proposed method is a feasible way of classifying different cancer types in applications.

References

[1] A.A. Alizadeh et al., "Distinct Types of Diffuse Large B-cellLymphoma Identified by Gene Expression Profiling," Nature,vol. 403, pp. 503-511, Feb. 2000

[2] T.Y. Lin and N. Cercone, Rough Sets and Data Mining: Analysis forImprecise Data. Springer, 1997

[3] J.M. Wei, S.Q. Wang, M.Y. Wang, J.P. You, and D.Y. Liu, "RoughSet Based Approach for Inducing Decision Trees," Knowledge-Based Systems, vol. 20, no. 8, pp. 695-702, Dec. 2007.

[4] Z. Pawlak, S.K.M. Wang, and W. Ziarko, "Rough Sets: Probabilisticversus Deterministic Approach," Int'l J. Man-Machine Studies,vol. 29, pp. 81-95, 1988.

[5] T.Y. Lin and N. Cercone, Rough Sets and Data Mining: Analysis forImprecise Data. Springer, 1997.

[6] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R.Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C.Cheng, D. Campana, D. Wikins, X. Zhou, J. Li, H. Liu, C.H. Pui,W.E. Evans, C. Naeve, L. Wong, and J.R. Downing, "Classification,Subtype Discovery, and Prediction of Outcome in Pediatric AcuteLymphoblastic Leukemia by Gene Expression Profiling," CancerCell, vol. 1, no. 2, pp. 133-143, 2002