Indian J. Innovations Dev., Vol. 1, No. 5 (May 2012) ISSN 2277 - 5390

Rough set theory based attribute reduction for breast cancer diagnosis

Sridevi T[1*], Murugan A[2]

[1]*Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India, Email: sripavi_mat@yahoo.com*

[2]*Department of Computer Science, Dr. Ambedkar Govt. Arts College, Chennai, Tamil Nadu, India*

[*]***Corresponding author**:* Research Scholar, Mother Teresa Women's University, Kodaikanal. Tamil Nadu, India, Email: sripavi_mat@yahoo.com

Abstract

Data mining (DM) techniques are used to determine interesting patterns from different domains according to the need of applications and the analyst. Medical field is one among the major user of the mining technology for diagnosing the attributes for the medical issues. Breast cancer is one of the most important medical problems. The modern researchers and technological advancements attempted to determine the cause and prevention in an effective manner with lesser number of attributes. But the diagnosis is lengthy process with multiple and multilevel attribute analysis in certain cases. In order to improve the accuracy of diagnosis with limited attributes, in this paper rough set based relative reduct algorithm is used to reduce the number of attributes using equivalence relation. The effectiveness of proposed Rough Set Reduction algorithm is analyzed on Wisconsin Breast Cancer Dataset (WBCD) and presented as a part of the paper. The experimental results show that the relative reduct performs better attribute reduction.

***Keywords*:** Data mining; Data Preprocessing; Rough Set; Data reduction; Breast Cancer Diagnosis.

***Abbreviations*:** DM-Data mining; WBCD-Wisconsin breast cancer dataset; KDD- Knowledge discovery in databases; RST-Rough sets theory;

Introduction

Data mining is the process of selecting, exploring and modeling large amounts of data to discover previously unknown patterns (Agrawal *et al*., 1996). Data mining treats as synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. "Data preprocessing" an important step in the knowledge discovery process, can be even considered as a fundamental building block of data mining. It enables data mining algorithms to be adopted easily to improve the effectiveness and the performance of the mining process.

In the mining process, the result and the patterns are based on the preprocessed data set. Attribute reduction is one of the important and frequently used techniques in data preprocessing for data mining (Guyon and Elissee, 2003). It deals in finding minimal subset of original attribute set by eliminating redundant attributes while maintaining the information of the problem in hand. Rough set theory can be used as such a tool to discover data dependencies and to reduce the number of attributes contained in a data set using the data alone and no additional information (Liu and Motoda, 1998). It is an extension of set theory for study of the intelligent systems characterized by insufficient and incomplete information. For medical diagnosis, data reduction is a critical problem, because it often contains an enormous quantity of data. The computation of the reduct from a rough set decision table is a way of selecting relevant features.

Methodology

Rough set theory

Rough sets theory (RST) is a mathematical tool for data analysis. It does not need external parameter to analyze and make conclusions about the datasets. It is new data mining method for study data integrity, knowledge uncertainty proposed by Pawlak (1982). Given a dataset with discretized

Indian J. Innovations Dev., Vol. 1, No. 5 (May 2012) ISSN 2277 - 5390

attributes, it is possible to find a reduct of original attributes that are most predictive of the class attribute. Rough set reducts can be found by using degree of dependency or using discernibility matrix.

Rough sets offer many opportunities for developing many knowledge discovery methods using partition properties and discernability matrix (Guyon and Elissee, 2003). Rough sets have many applications in KDD among them; feature selection, data reduction, and discretization are frequently used techniques (Zdzislaw Pawlak, 1991; Qiang Shen and Alexios Chouchoulas, 2000). Rough set can be used to find subsets of relevant (indispensable) features (Jensen *et al*., 2001). Combining rough sets theory with a known classifier yields a filter feature selection method, since it uses the class label information to create the indiscernability relation. It provides a mathematical tool that can be used to find out all possible feature subsets.

In the process of RST, first a decision table containing object ids, the discretized attributes and the decision attribute are created. The class attribute of the data set has been considered as the decision attribute. Then the reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. The reduced set of attributes obtained by applying RST on the discretized data set has shown in the Table 1.

For an information System I = <U, A, V, F>, U ={x1,x2,…….xn} is a non-empty set of finite objects (the universe of discourse), A is a finite set of attributes {a1,a2,….an}, which can be further divided into two disjoint subsets of C and D, A={C ∪ D }where C is condition attributes and D is a set of decision attributes. $V = \bigcup_{a\in A} V_a$ and V_a is a domain of the attribute a, and F:U×A→V is the total decision function called the information function such that $F(x,a)\in V_a$ for every a∈A, x∈U.

For every set of attributes $P \subseteq A$, an indiscernibility relation IND(P) is defined in the following way: two objects x and y are indiscernible by the set of attributes P∈A if and only if $f(x, q) = f(y, q)\ \forall q \in P$. The equivalence class of IND(P) is called elementary set in P because it represents the smallest discernible groups of objects. For any element x of U, the equivalence class of $x \in IND(P)$ is represented as $[x]_P$.

A rough set approximates traditional sets using a pair of sets named the lower and upper approximation of the set. The lower and upper approximations of a set $P \subseteq U$, are defined as

$$\underline{P}(X) = \{x \in U \mid [x]_P \subseteq X\}$$

$$P(X) = \{x \in \bar{U} \mid [x]_P \cap X \neq \emptyset\}$$

The boundary region is defined as:

$$BND_P(X) = \underline{P}(X) - P(X)$$

It consists of those objects that can neither be ruled in nor ruled out as members of the target set *X*. The set is said to be rough if its boundary region is non-empty, otherwise the set is crisp. Assuming *P* and *Q* are equivalence relations in *U*, the important concept positive region $POS_P(Q)$ is defined as:

$$POS_P(Q) = U_{X\in U|Q}\ \underline{P}(X)$$

A positive region contains all objects of *U* that can be classified to classes of U/Q using the information in attributes P.

There often exist some condition attributes that do not provide any additional information about the objects in U in the information system. So, these redundant attributes can be eliminated without losing essential information. A reduct attribute set is a minimal set of attributes from A that provided that the object classification is the same as with the full set of attributes. Given C and D⊆A, a reduct is a minimal set of attributes such that

IND(C) = IND(D)

This reduction approach is implemented with the associative and relative reduct approach. In this

Indian J. Innovations Dev., Vol. 1, No. 5 (May 2012) ISSN 2277 - 5390

work adopted and tested with the relative reduct approach of the classification of the given data set. It is further divided into two groups and its result is integrated.

Data collection and description

WBCD taken from UCI machine learning repository (Blake *et al*., 1998) is considered for this study. The dataset contains 569 instances taken from needle aspirates from patients' breasts where 357 cases belong to benign class and 212 cases belong to malignant class. The descriptive attributes are recorded with four significant digits and include the nuclear radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean, the 3 extreme (usually the mean of the three largest values), and the standard error of each feature across the nuclei are obtained, resulting in a total of 30 variables. There is no missing value.

According to Street Wolberg *et al*. (1993), the 10 extracted attributes are: 1. Radius: obtained by averaging the length of radial line segments from the centroid to the individual snake points, 2. Perimeter: Sum over the total distance of the snake points, 3. Area: Number of pixels in the interior of the snake and adding 1/2 of the perimeter pixels, 4. Compactness: perimeter2/area, 5. Smoothness: Difference between length of a radial line and the mean length of the lines surrounding it, 6. Concavity: Draw chords between non-adjacent snake points and measure distance to object boundary, 7. Concave Points: Counts the number of contour concavities, 8. Symmetry: Similar to relation between major and minor axis, 9. Fractal Dimension approximated using the "coastline approximation" described by Mandelbroat *et al*. (1965), 10. Texture Variance of the intensity levels in the interior of the snake.

Results and Discussion

Implementation approach

All 30 attributes are represented in real valued measurement but for the purpose of rough set theory must be descritized, producing a new dataset with crisp values. The mean and standard deviations of each of these 30 attributes are separated by benign and malignant cases given in (Dash and Liu, 1997). Using this information, values of all 30 attributes in each record are replaced by either 0 or 1 according to mean.

In Han *et al*. (2004), relative reduct, a feature selection method based on an alternative dependency measure is presented. The technique is originally proposed to avoid the calculation of discernability functions or positive regions, which can be computationally expensive without optimizations. The approach attempted with traditional rough set degree of dependency with an alternative measure (Jensen, 2004) selects more optimal number of features perfectly, since it performs backward elimination process. The degree of dependency is defined as follows:

$$K_R(D) = \frac{|U/IND(R)|}{|U/IND(R \cup D)|}$$

The relative reduct (RR) algorithm.

RR (*C*, *D*)

C, the set of all conditional features;

D, the set of decision features;

(1) $R \leftarrow C$

(2) $\forall a \in C$

(3 if $\kappa_{R-\{a\}}(D)=1$

(4) R←R{a}

(5) return *R*

Using backward elimination, attributes are removed from the set of considered attributes if the relative dependency equals 1. Upon their removal attributes are consider one at a time, starting with the first and evaluating their relative dependency.

From the selected data set the initial five attributes are adopted for the reduction process and named as initial set I .I={R,T,P,A,S)} are conditional attributes and {D} is a decision attribute (Table 1).

Indian J. Innovations Dev., Vol. 1, No. 5 (May 2012) ISSN 2277 - 5390

Where R – Radius, T-Texture, P-Perimeter, A-Area and S- Smoothness, Decisions attribute values: B – benign, M- Malignant.

Using backward elimination algorithm first the attribute R is considered for elimination:

$$K_{\{T,P,A,S\}}(D) = \frac{\frac{U}{IND(T,P,A,S)}}{\left|\frac{IND(T,P,A,S,D)}{U}\right|}$$

$$\frac{\{1\}\{2,3,7\}\{4,5,8\}\{6,9\}\{10\}}{\{1\}\{2,3,7\}\{4,5,8\}\{6,9\}\{10\}} = \frac{5}{5}$$

As the relative dependency is equal to 1, attribute R can be removed from the reduct. Hence the current reduct is {T, P, A, S}. The algorithm

Table 1. reduced set of attributes obtained by applying RST on the discretized data set

x∈U	R	T	P	A	S	D
1	0	0	0	0	1	B
2	0	1	0	0	0	B
3	0	1	0	0	0	B
4	0	0	0	0	0	B
5	0	0	0	0	0	B
6	1	0	1	1	1	M
7	0	1	0	0	0	B
8	0	0	0	0	0	B
9	1	0	1	1	1	M
10	0	1	0	0	1	M

then considers the elimination of attribute T, as the relative dependency is not equal to 1, attribute T is not removed. The algorithm then evaluates the elimination of attribute P from reduct, further processed and reached with dependency is not equal to 1, hence attribute P is retained in the reduct. Similarly, A and S are consider for elimination. Finally the current reduct is {T, P, and S}. As there is no further attribute to consider, the algorithm terminates and outputs the reduct {T, P, S}. The obtained results for two groups are presented below.

Analysis of result

The dataset is implemented with the attribute reduction process. Out of all three set of values it is grouped into the basic elements, standard deviation error and the worst case elements. While evaluating the values, the results are attempted to reduce the basic elements and the order is given in Table 2. The values are reduced using relative reduct and obtained the result of ten attributes to seven attributes. The results are obtained at the same level. The same set of standard deviation values are tested and obtained the same result.

Conclusion

Table 2. results of reduce attempted basic elements

Selected Group	Reduced Elements	Set Elements
V1,2,3,4,5	1,2,3,5	34
V6,7,8,9,10	6,8,9,10	34
1,2,3,5	Not able to reduct	
6,8,9,10	8,9,10	12
1,2,3,4,5,6,7,8,9,10	1,2,3,5,8,9,10	Integrated

The obtained relative reduct algorithm selects the attributes which could be eliminated from the divided set attributes of the whole set. The attribute set and its similar properties are divided into two and the reduction process achieved in the iterative process. While processing the attribute reduction it reached the maximum level. The same result is obtained for the similar property. Therefore it could finalize such a way that the observed basic elements values and its standard deviation error results are provided in the same impact level of the obtained result and the reducted attributes. Now the ten attributes are reduced and reached into seven after two iteration. If the process leads for the set of 30 attributes, the result could be obtained with twenty one attributes instead of 30 attributes. Therefore this algorithmic approach is appreciated for the attribute reduction. The reduction process with the real time applications must be validated with large heterogeneous data set to ensure the adopted techniques.

Acknowledgement

To the maintainers of the UCI repository of machine learning databases.

References

Indian J. Innovations Dev., Vol. 1, No. 5 (May 2012) ISSN 2277 - 5390

1. Agrawal R, Imielinski T and Swami A (1993) Database mining: A performance perspective. *IEEE Trans. Knowl. Data Eng.* 5(6), 914–925.
2. Blake CL and Merz CJ (1998) UCI Repository of machine learning databases, Irvine, University of California, http://www.ics.uci.edu/~mlearn/
3. Dash M and Liu H (1997) Feature Selection for Classification. *Intell. Data anal.* 1(3), 131-156.
4. Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R (1996) *Advances in Knowledge Discovery and Data Mining*, pages 495–515. AAAI Press / the MIT Press.
5. Guyon I and Elissee A (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157-1182.
6. Han J, Hu X and Lin TY (2004) Feature subset selection based on relative dependency between attributes, in *Proc. of the 4th International Conf. on Rough sets and Current Trend in Computing,* Uppsala, pp. 176–185.
7. Jensen R and Shen Q (2001) A Rough Set-Aided System for Sorting WWW Bookmarks, In Zhong N *et al*. (Eds.), Web Intelligence: Research and Development, pp. 95-105.
8. Jensen R (2004) Combining rough and fuzzy sets for feature selection, Ph.D thesis, University of Edinburgh.
9. Liu H and Motoda H (1998) Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers.
10. Mandelbrot BB (1965) Linear and nonlinear separation of patterns by linear programming. *Oper. Res.* 13, 444-452.
11. Qiang Shen and Alexios Chouchoulas (2000) A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. *Eng. Appl. Artif. Intell.* 13(3), 263–278.
12. Quinlan, JR (1993) C4.5: Programs for Machine Learning, The Morgan Kaufmann Series in Machine earning. Morgan Kaufmann Publishers, San Mateo, CA.
13. Street W, Wolberg W and Mangasarian O (1993) Nuclear feature extraction for breast tumor diagnosis. Available from: citeseer.ist.psu.edu/street93nuclear.html.
14. Zdzislaw Pawlak (1982) Rough sets. *Int. J. Compu.Info. Sci.* 11, 341-356.
15. Zdzislaw Pawlak (1991) *Rough Sets-Theoretical Aspects and Reasoning about Data*, Kluwer Academic Publications.