Sridevi T, Murugan A / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 3, May-Jun 2013, pp.772-776

An Expert System To Detect Malignant Cells In Breast Cancer

Sridevi T[1], Murugan A[2]
[1]Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamil Nadu,India
[2]Department of Computer Science, Dr. Ambedkar Govt. Arts College, Chennai, Tamil Nadu, India

Abstract

Data mining is the process of analyzing vast amount of data and extracting useful information. In medical diagnosis, data mining techniques have been used to discover hidden relationships and trends i.e. valuable knowledge.In order to achieve successful data mining, feature selection is an indispensablecomponent. It is a process of selecting a subset of original features according to certain criteria, andan important and frequently used technique in data mining for dimension reduction. Rough set has been one of the most successful methods used for medical feature selection.Breast cancer is one of leading causes of death among women in worldwide countries, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. This paper presents an automatic system for detection of cancerous cells in breast cancer using rough set theory.

Keywords:*Data mining, Rough set, Feature selection, Breast cancer Diagnosis, Classification*.

1. INTRODUCTION

Data mining is an interdisciplinary research area such as machine learning, intelligent information systems, statistics,database systems and expert systems. Hence data mining has become a research area with increasing importance. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. Classification is an important data mining tasks to classify the data in the process of Knowledge Data Discover as they try to find meaningful ways to interpret data sets. Classification is one of the data mining and machine learning problem focusing great attention recently in the database community [2].

Diagnosis and medical issues are important applications of classification. Breast cancer diagnosis has been a challenging research problem for many researchers in the present decades. Medical data often contains a huge number of irrelevant and redundant features and a relatively small number of cases, which dramatically impact quality of disease diagnosis. As a result, feature selection is expected to improve differentiation performance. Feature selection is a process which attempts to select more relevant features. Rough set theory (RST) has been recognized to be one of the powerful tools in the medical feature selection. The computation of the core and reduct from a rough set decision table is a way of selecting relevant features [3]. Mining on a reduced set of features has an additional benefit. It reduces the number of features appearing in the discovered patterns, helping to make the patterns easier to understand.

In this paper we use feature selection algorithm based on rough set theory and investigate the effectiveness of the method on the breast cancer diagnosis dataset of UCI machine learning repository.

2. ROUGH SET

Rough set theory was introduced by Pawlak in 1982. It was developed based on mathematical tool to deal with vagueness and uncertainty in the classification of objects in a set [4]. Rough set theory is a good candidate for classification applications. Various efforts have been made to improve the efficiency and effectiveness of classification with rough sets [5]. It does not need external parameter to analyze and make conclusion about the datasets. The rough set philosophy is founded on the assumption that there is some information regarding features which can be associated with every object of the universe. In rough sets, the data is organized in a table called decision table, which are flat tables containing attributes as columns and data elements as rows. The class label is called as decision attribute, the rest of the attributes are the condition attributes. 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 sample of 10 records of breast cancer diagnosis dataset has shown in the Table2.

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 $\cup$ D }where C is condition attributes and D is a set of decision attributes. V = $\cup_{a \in A} V_a$ 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)∈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 $[x]_P$.Rough set theory defines three regions based on the equivalent classes induces by the attribute values: lower approximation, upper approximation, and boundary.

The lower and upper approximations of a set $P \subseteq U$, are defined as

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

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

The boundary region is defined as:

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

It consists of those objects that can neither be ruled in nor ruledout 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:

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

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 thatIND(C) = IND (D).

3. METHODOLOGY

Table 1: The detail of the nine attributes of breast cancer data

Label	Attribute	Domain
A	Clump Thickness	1-10
B	Uniformity of Cell Size	1-10
C	Uniformity of Cell Shape	1-10
D	Marginal Adhesion	1-10
E	Single Epithelial Cell Size	1-10
F	Bare Nuclei	1-10
G	Bland Chromatin	1-10
H	Normal Nucleoli	1-10
I	Mitoses	1-10

3.1 Dataset Description

Breast cancer diagnosis data set has been taken from UCI machine learning repository [6]. The dataset contains 699 instances taken from needle aspirates from patients' breasts, where 16 instances have missing values, these are replaced by mean value, of which 458 cases belong to benign class and the remaining 239 cases belong to malignant class. Each record in the database has nine attributes. The nine attributes are listed in Table 1 are graded 1-10, with class attribute represented as 2 for benign and 4 for malignant cases.

Breast cancer is an uncontrolled growth of breast cells. A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous).

Table 2: Sample instances of Breast Cancer data

x∈U	A	B	C	D	E	F	G	H	I	CLASS
1	4	2	1	1	2	1	2	1	1	2
2	1	1	1	1	1	1	3	1	1	2
3	2	1	1	1	2	1	2	1	1	2
4	5	3	3	3	2	3	4	4	1	4
5	1	1	1	1	2	3	3	1	1	2
6	8	7	5	10	7	9	5	5	4	4
7	7	4	6	4	6	1	4	3	1	4
8	4	1	1	1	2	1	2	1	1	2
9	4	1	1	1	2	1	3	1	1	2
10	10	7	7	6	4	10	4	1	2	4

3.2 Rough set based feature selection algorithm

Our breast cancer diagnosis classifier is based on the concept of rough set theory using forward selection. It attempts to calculate a minimal reduct without exhaustively generating all possible subsets [7]. The problem or finding minimal reduct of an information system has been the subject of much research [8]. It starts with an empty set of attributes. The best of the original attributes is determine and added to the set using the dependency $\gamma_c(D) = POS_c(D) / U$where U is the cardinality of set U, $POS_c(D)$ called positive

region, is defined by $POS_c(D) = U_{x \in U/D}(x)$. At each subsequent iteration or step, the best (i.e. greatest increase in dependency) of the remaining original attributes is added to the set until the dependency of the reduct candidate equals the consistency of the dataset (1 if the dataset is consistent).

The reduction of attributes is achieved by comparing equivalence relation generated by sets of attributes. Attributes are removed so that the reduced set provides the same quality of classification as the original.

To evaluate the performance of the algorithm, we use classification accuracy of classifier. In order to make the observation more convincing, in this work, we use three phases with different number of instances.To evaluate the effectiveness of rough set based feature selection algorithm, we attempt to compare the classification accuracy of various classifier for three different record sets with instances 100,350 and the whole record set 699. Our reported accuracies are the mean of the ten accuracies from ten-fold cross-validation. This technique ensures that the training and test sets are disjoint. In this paper BayesNet, NaiveBayes, RBFNetwork, MLP, SMO, IBK and J48 classifiers are used to compute the classification accuracies for three phases.

4. EXPERIMENTAL ANALYSIS

The rough set based attribute reduction algorithm has been implemented using MATLAB. It is used to eliminate the unimportant and redundant features. The reduced attribute set obtained for Breast cancer dataset with 100 instances is: {Uniformity of Cell Size, Uniformity of cell Shape, Bland Chromatin}. Breast cancer dataset with 350 instances is:{Clump Thickness,Uniformity of cell Shape, Marginal Adhesion, Bare Nuclei } and Breast cancer dataset with all 699 instances is :{Clump Thickness, Uniformity of cell Shape, Bare Nuclei, Bland Chromatin}.In this paper WEKA toolkit is used to analyze the dataset with the data mining algorithms[9]. The toolkit is developed in Java and is open source software issued under the GNU General public License [10].

Table 4: Comparison Results for three different numbers of instances

Breast cancer dataset (BCD)	Instances	No. of attributes	No. of reduct
BCD 1-100	100	9	3
BCD 1-350	350	9	4
BCD 1-699	699	9	4

Table 5: Classification accuracy for breast cancer data setwith reduced set of three different numbers of instances

S.No.	Algorithm	100 instances	350 instances	699 instances
1	BN	92	94.8571	96.8526
2	NB	87	95.1429	95.8512
3	RBF	91	94	95.9943
4	MLP	93	95.7143	96.2804
5	SMO	91	95.1429	96.2804
6	JBK	91	93.4286	95.9943
7	J48	92	93.4286	94.9928

Figure 1: Performance analysis of the three differentnumbers of instances with reduced attribute set

Root Mean-Squared Error: The root mean square error calculates the differences between values predicted by a model and the values actually observed. It is used to measure the accuracy. It is ideal if it is small.

Table 6: Comparison of Root Mean-Squared Error

S.No.	Classifier	100 instances	350 instances	699 instances
1	BN	0.2592	0.2143	0.1683
2	NB	0.2961	0.2048	0.1909
3	RBF	0.2868	0.2142	0.1814
4	MLP	0.2312	0.2159	0.1811
5	SMO	0.3	0.2204	0.1929
6	JBK	0.2712	0.2528	0.1953
7	J48	0.2702	0.2423	0.2107

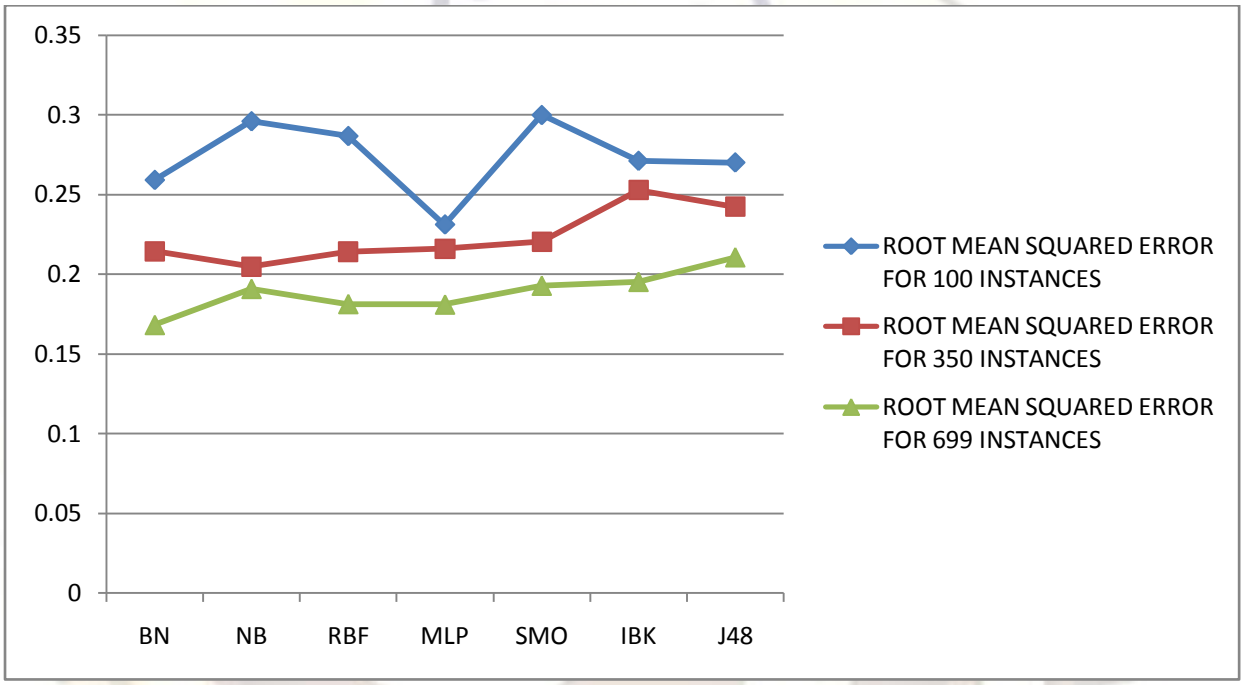

Figure 2:Comparison of root mean squared error

The data set is implemented with the attribute reduction process. Out of all nine attributes, after evaluating the values the attribute set is reduced by {B,C,G} for 100 instances , {A,C,D,G} for 350 and (A,B,F,G} for all 699 instances. Based on classification techniques using 10 fold cross validation, it is clear that all classifiers show the higher accuracy in increasing number of instances.When the number of instances increases, through study the classifiers have the rich "experience", the error rate drop, and when instances increases to a certain degree, the number of attributes in reduced set is stability. It is depicted in Table 4 and Table 6. and indicates that the proposed algorithm is effective and efficient, especially for the large data sets.

CONCLUSION

In this paper, seven different classifiers are used for the classification of data. These techniques are applied on breast cancer diagnosis data set with different numbers of instances. The fundamental concept to take different dimensionality is to analyze the performance of the discussed feature selection for small as well as large dataset. On the basis of comparison done over accuracy, the classifications with highest accuracy are obtained for large data sets. The experimental results show that the rough set based attribute reduction performs better attribute reduction on large data sets. Our approach shows an excellent performance, not only high classification accuracy, but also with respect to the number of features selected.

ACKNOWLEDGEMENT

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

REFERENCES

[1]Han and M.Kamber, *"Data Mining : Concepts and Techniques, "*Morgan Kaufmann. 2000

[2] Dr. DSVGK Kaladhar, B.Chandana and P.Bharath kumar, *"Predicting cancer survivability using classification algorithms,"* IJRRCS,2(2),pp. 34-343. 2011.

[3] D. SleZak, *"Variuous approaches to reasoning with frequency-based decision reducts: a survey,"* In L.Polkowski, S.Tsumoto and T.Y.Lin, editors, Rough sets in Soft computing and knowledge discovery : New Development physica verlag. 2000.

[4] Zdzislaw Pawlak "*Rough Sets-Theoretical Aspects and Reasoning about Data",* Klower Academic Publication. 1991.

[5] N.Zhong, J.Z. Dong and S.Ohsuga, *"Using Rough Sets with Heuristics for feature selection,"* Journal of Intelligent Information Systems,16.199-214,2001.

[6] ww.ics.uci.edu/~mlearn.

[7] A.E.Hassanien, Z.Suraj, D.Slezak, and P.Lingras, *"Rough Computing: Theories, Technologies, and Applications,"* NewYork: Information Science Reference, 2008.

[8] J.J. Alpigini, J.F. Peters, J.Skowronek, N. Shong(Eds.): *"Rough sets and Current Trends in Computing",* Third International Conference, RSCTC 2002. Malvern,PA,USA,October 14-16,2002. Lecture Notes in Computer Science 2475 Springer 2002, ISBN 3-540-44274-X.

[9] Weka: Data Mining Software in java http://www.cs.waikato.ac.nz/ml/weka/

[10] Ian H.Witten and Elbe Frank, "*Datamining Practical Machine Learning Tools and Techniques,"* Second Edition,Morgan Kaufamann, San Fransisco. 2005.