ISSN:2320-0693
Vol-19-Issue-24-June-2020

Performance Analysis of Clustering and Data Mining Methods on Medical and Clinical Data Through Rough Set Theory

Mr.Hiren R. Kavathiya
Assistant Professor
Department of C.S.& I.T.
Shree M.V. & Smt. N.Virani Science College, Rajkot
Email:hrkavathiya@vsc.edu.in

Dr. G. C.Bhimani
Professor & Head,
Department of Statistics
Saurashtra University, Rajkot
Email:gcbhimani@yahoo.co.in

Abstract:
In this paper, we used two models for analysis, prediction, cost estimation, cost sensitivity and knowledge discovery rules for medical data by using rough set theory. In the first model,RGA and rule generation algorithms where used within which concepts of Generic Algorithms were utilized for the optimization in the analysis for the better performance and error reduction. Various attributes like scaling , itching, polygonal, age, erythematic etc were considered and and its count and percentage were calculated based on the model. Rules were distributed in various classes and its count was measured. Three methods were called for classification result generation viz. direct method, indirect method and meta-cost sensitive method and the result found were upto the mark. In the second method, Genetic, Exhaustive, Covering and LEM2algorithm were used and specificaaly they emphasized on PIMA data set for the analysis and rule generation.

Keyword: Data Mining, Genetic Algorithm, Optimization, Rough Set Theory

Introduction:

A huge amount of data is stored in clinical databases about the patient's diagnosis, lab-test results, patients' treatments etc., and is a gold mine of medical knowledge for medical researchers and doctors. The study of medical data by using the Data Mining (DM) techniques is an unmapped frontier which needs extra attention and hence, several kinds of medical data is taken into consideration for the present investigation on medical mining. This sudden rise of data needs an automated way to source valuable information and therefore, it can be concluded that data mining can be applied to the medical domain. Crucial information can be extracted through data mining. Also, the regularities and the revealed data can be later applied in the corresponding field to upsurge the working efficacy and improve the quality of decision making. Data mining is an important step in the knowledge discovery process and it deals with the issue of finding regularities and patterns in data. Medical informatics is the field of information science related to the analysis and dissemination of medical data via application of computers to several facets of health care plus medicine. Here, we focus on applications of Data mining techniques in Hospital Information Systems (HIS) such as emergencies, ward management, human resources, physical resources, etc., for improving the quality. While large amount of data is available, the efficacy is not adequate. Therefore, a detailed investigation will be made as to know how the DM techniques can be implemented retrospectively on massive data in an automated manner by keeping in mind the quality improvement facet. As a matter of fact, using DM, organizations can upturn the productivity of their dealings with customers, detect deception and improve risk management. In many developed countries, routine health tests are common and are widespread among the adult or grown population. Hence, cost effective precautionary measures and the detection of disease in its initial stages of development, give patients a better opportunity of treatment as compared to a later stage. In any case that the healing of the patient is doubtful or unknown, the treatment is still helpful and provides life comfort.

The main focus in this paper is on various ways of using the Rough Set Theory

Model 1:

Cost Sensitivity Analysis and the Prediction of Optimal Rules for Medical Data by using Rough Set Theory

The following is investigated by this model.

(i) Reducts as well as the rules generated by the rough set approach.

(ii) Rules generation by direct and indirect methods.

(iii) The generation of optimal rules for cost effectiveness associated with the dermatology data set.

(iv) The reducts of the data are found by applying the rough set reduction technique as they contain minimal subset of attributes associated with class label for the purpose of prediction. A reduct can be defined as a minimal set of attributes such that B ⊆ A in a way that IND(B) = IND(A); IND(X) is called the X-indiscernibility relation. All the generated reducts also generate the rough sets dependencies.

Algorithm 1- RGA

```
Input: Information table (IFT) consisting discretized real valued attribute.
Output: Reduct set Rf= {r1U r2 U……………rn}
Step 1: for each condition attributes c in C do
Compute the correlation factor(CF) between c and the decisions attributes D
Step 2: if CF >0 then
Step 3: c -->relevant attributes
Step 4: end if
Step 5: end for
Step6:Partition the set of relevant attributes into different variable sets
Step 7: for each variable sets do Compute the dependency degree (DD) and classification accuracy(CA)
Step 8: Initialize the set with highest CA and DD as the reductset(Rinit)
Step 9: end for
Step 10: for each r in R do Compute DD between D and Rf
Step 11: Merge the attributes produced in previous step with the rest of conditional attribute
Step12:Compute the discrimination factor(DF) for each combination to find the highest DFmax
Step13: Add the combinations with DFmax to Rf
Step 14: end for
Step 15: repeat step 10
Step 16: until all the attributes Rinit are processed
```

Algorithm 2: Rule Generation

```
Input: Universal set U, Reduct sets Rf= r1 U r2 U :::: U rng
Output: Set of rules
Step 1: for each reductr do
Step 2: for each correspondence object x do
Step 3: Contract the decision rule (c1 = v1 ^ c2 = v2 ^ :::: ^ cn= vn) →d = u
Step 4: Scan the reductr over an object x
Step 5: for every c in C do
Step 6: Assign the value v to the correspondence attribute a
Step 7: end for
Step 8: Construct a decision attribute d
Step 9: Assign the value u to the correspondence decision attribute d
Step 10: end for
Step 11: end for
```

Result:

The attribute occurrence of reducts is represented in table 1 while the number of reducts for every attribute is shown above. The reduct number is found to be 10 while 1 is the size of the core.

Table 1 : Occurrence of attributes in Reducts

Attrib	Count	Percent	Core
Scaling	2	3	No
Itching	8	11.9	No
Polygonal	3	4.5	No
Exocytosis	3	4.5	No
Inflammatory	9	13.4	No
Age	10	14.9	Yes
Erythema	4	6	No
parakeratosis	4	6	No
Elongation	3	4.5	No
Knee	1	1.5	No
Focal	4	6	No
Spongiosis	5	7.5	No
Acanthosis	5	7.5	No

Definite	2	3	No
Koebner	1	1.5	No
Hyperkeratosis	2	3	No
PNL	1	1.5	No

Table 2: Reducts within Each Attribute

(1-10)	Size	Pos.Reg.	SC	Reducts
1	6	1	1	{scaling, itching, polygonal, exocytosis, inflammatory, Age}
2	6	1	1	{erythema, itching, parakeratosis, elongation, inflammatory, Age}
3	7	1	1	{erythema, itching, knee, focal, spongiosis, inflammatory, Age}
4	7	1	1	{erythema, scaling, itching, polygonal, acanthosis, elongation, Age}
5	7	1	1	{erythema, acanthosis, parakeratosis, focal, spongiosis, inflammatory, Age}
6	6	1	1	{define, itching, exocytosis, acanthosis, inflammatory, Age}
7	7	1	1	{definite, itching, koebner, focal, spongiosis, inflammatory, Age}
8	7	1	1	{itching yperkeratosis, parakeratosis, elongation, spongiosis, inflammatory, Age}
9	7	1	1	{itching, polygonal, PNL, exocytosis, acanthosis, inflammatory, Age}
10	7	1	1	{acanthosis, hyperkeratosis, parakeratosis, focal, spongiosis,inflammatory, Age}

The reducts generate a total of 3550 rules. The distribution of rules is shown in table 5.2 for a number of classes. By using genetic algorithm, the rules generated from the reducts are presented in fig. 5.3 and 5.4 and maximum number of them is for psoriasis class. 42 rules are formed by covering algorithm. The maximum number of counts as noticed from table 5.3 is lichen while the accuracy and coverage for diverse classes is predicted in table 5.6a which are the decision classes' unity.

Table 3 : Distribution of rules among classes

Decision Class	**Count**
Seboreic	608
Psoriasis	1047
Lichen	709
Cronic	515
Pityriasis	473
Rubra	198

Table 4: Distribution of Decision rules for each class

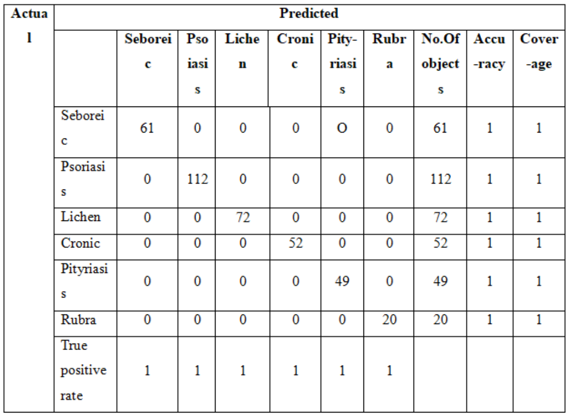
(1-3555)	Size	Reducts
1	1	(scaling=2)&(itching=3)&(polygonal=0)&(exocytosis=3)&(inflammatory=1)&(Age=55)=>(CLASS={seboreic[1]})
2	1	(scaling=3)&(itching=2)&(polygonal=0)&(exocytosis=1)&(inflammatory=1)&(Age=8)=>(CLASS={psoriasis[1]})
3	1	(scaling=1)&(itching=3)&(polygonal=3)&(exocytosis=1)&(inflammatory=2)&(Age=26)=>(CLASS={lichen[1]})
4	1	(scaling=2)&(itching=0)&(polygonal=0)&(exocytosis=0)&(inflammatory=3)&(Age=40)=>(CLASS={psoriasis[1]})
5	1	(scaling=3)&(itching=2)&(polygonal=2)&(exocytosis=1)&(inflammatory=2)&(

		Age=45)=>(CLASS={lichen[1]})
6	1	(scaling=3)&(itching=0)&(polygonal=0)&(exocytosis=2)&(inflammatory=1)&(Age=41)=>(CLASS={seboreic[1]})
7	2	(scaling=1)&(itching=2)&(polygonal=0)&(exocytosis=1)&(inflammatory=2)&(Age=18)=>(CLASS={cronic[2]})
8	1	(scaling=2)&(itching=3)&(polygonal=3)&(exocytosis=2)&(inflammatory=3)&(Age=57)=>(CLASS={lichen[1]})
9	2	(scaling=2)&(itching=0)&(polygonal=0)&(exocytosis=2)&(inflammatory=2)&(Age=22)=>(CLASS={pityriasis[2]})
10	1	(scaling=2)&(itching=0)&(polygonal=0)&(exocytosis=3)&(inflammatory=2)&(Age=30)=>(CLASS={pityriasis[1]})
11	1	(scaling=3)&(itching=1)&(polygonal=0)&(exocytosis=0)&(inflammatory=1)&(Age=20)=>(CLASS={psoriasis[1]})
12	1	(scaling=2)&(itching=3)&(polygonal=0)&(exocytosis=2)&(inflammatory=1)&(Age=21)=>(CLASS={seboreic[1]})
13	1	(scaling=3)&(itching=2)&(polygonal=0)&(exocytosis=3)&(inflammatory=1)&(Age=22)=>(CLASS={seboreic[1]})
14	1	(scaling=3)&(itching=0)&(polygonal=0)&(exocytosis=0)&(inflammatory=2)&(Age=10)=>(CLASS={psoriasis[1]})
15	1	(scaling=2)&(itching=3)&(polygonal=3)&(exocytosis=1)&(inflammatory=1)&(Age=65)=>(CLASS={lichen[1]})
16	1	(scaling=1)&(itching=1)&(polygonal=0)&(exocytosis=1)&(inflammatory=2)&(Age=40)=>(CLASS={pityriasis[1]})
17	1	(scaling=2)&(itching=3)&(polygonal=0)&(exocytosis=2)&(inflammatory=1)&(Age=30)=>(CLASS={seboreic[1]})
18	1	(scaling=3)&(itching=0)&(polygonal=0)&(exocytosis=0)&(inflammatory=2)&(Age=38)=>(CLASS={psoriasis[1]})
19	1	(scaling=1)&(itching=3)&(polygonal=3)&(exocytosis=3)&(inflammatory=2)&(Age=23)=>(CLASS={lichen[1]})
20	1	(scaling=1)&(itching=3)&(polygonal=0)&(exocytosis=0)&(inflammatory=2)&(Age=17)=>(CLASS={cronic[1]})
21	1	(scaling=1)&(itching=2)&(polygonal=0)&(exocytosis=0)&(inflammatory=1)&(Age=8)=>(CLASS={rubra[1]})
22	1	(scaling=2)&(itching=0)&(polygonal=0)&(exocytosis=2)&(inflammatory=2)&(Age=51)=>(CLASS={seboreic[1]})
23	1	(scaling=2)&(itching=2)&(polygonal=0)&(exocytosis=1)&(inflammatory=2)&(Age=42)=>(CLASS={cronic[1]})
24	1	(scaling=2)&(itching=3)&(polygonal=2)&(exocytosis=0)&(inflammatory=2)&(Age=44)=>(CLASS={lichen[1]})
25	1	(scaling=0)&(itching=3)&(polygonal=0)&(exocytosis=2)&(inflammatory=2)&(Age=22)=>(CLASS={cronic[1]})

Table 5 : Distribution of rules among classes for 4 cases

Decision Class	Count
Psoriasis	12
Lichen	18
Cronic	5
Rubra	7

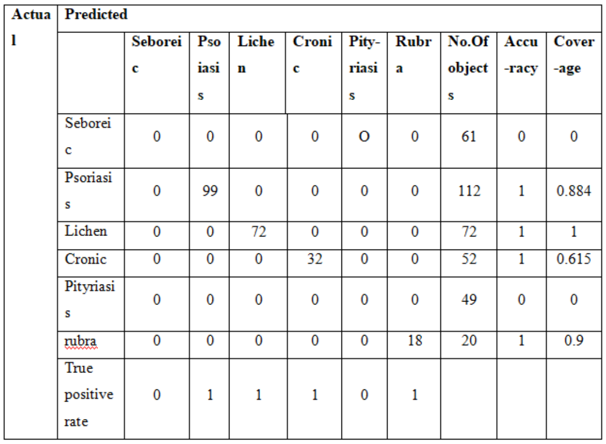
Table 6 a : Confusion matrix for rules generation by Genetic algorithm



Actual	Predicted								
	Seboreic	Psoiasis	Lichen	Cronic	Pity-riasis	Rubra	No.Of objects	Accu-racy	Cover-age
Seboreic	61	0	0	0	O	0	61	1	1
Psoriasis	0	112	0	0	0	0	112	1	1
Lichen	0	0	72	0	0	0	72	1	1
Cronic	0	0	0	52	0	0	52	1	1
Pityriasis	0	0	0	0	49	0	49	1	1
Rubra	0	0	0	0	0	20	20	1	1
True positive rate	1	1	1	1	1	1			

Table 6b Confusion matrix for rules generation by Covering algorithm

ISSN:2320-0693
Vol-19-Issue-24-June-2020



Actual	Predicted								
		Seboreic	**Psoiasis**	**Lichen**	**Cronic**	**Pityriasis**	**Rubra**	**No.Of objects**	**Accuracy**	**Coverage**
	Seboreic	0	0	0	0	O	0	61	0	0
	Psoriasis	0	99	0	0	0	0	112	1	0.884
	Lichen	0	0	72	0	0	0	72	1	1
	Cronic	0	0	0	32	0	0	52	1	0.615
	Pityriasis	0	0	0	0	0	0	49	0	0
	rubra	0	0	0	0	0	18	20	1	0.9
	True positive rate	0	1	1	1	0	1			

Table 7 : Classification results

Measures	Direct Method	Indirect Method	Meta-cost Sensitive
Correctly Classified	91.25%	95.90%	72.95%
Incorrectly classified	8.74%	4.09%	27.04%
Kappa	0.8903	0.9488	0.6634
Accuracy	91.3	96.9	69.4
ROC	94.8	97.5	84.3
Total Cost	123	106	119

Conclusion – Model 1

Every test performed in the medical world for the diagnosis of diseases is considered as a feature. Expensive and irrelevant tests can be avoided using the process of feature selection. Optimal subset of features is an important decision when processing medical data in order to improve the classification performance of the model built from the selected data. The main focus of this model is as follows.

(i) Rough set approach generated reducts and rules

(ii) Direct and indirect methods generated rules

(iii) Optimal rules are generated for the purpose of cost effectiveness associated with the dermatology data set.

The present investigation shows the following interesting features:

(i) Maximum classification accuracy is yielded by the indirect approach from all the three classifiers (Direct, Indirect and Meta classifiers).

(ii) A different structure of cost curves is produced. This prediction is done by the Cost sensitivity analysis along with the fact that the decision tree classifier exhibits the optimum cost.

(iii) Dimensionality reduction approach is used to present optimal set of rules by the Rough set approach wherein GA and Covering are the algorithms used.

Model 2: Rough Set Approach for Novel Decision Making in Medical Data for Rule Generation and Cost Sensitiveness

Genetic Algorithm

```
Choose initial population Evaluate the fitness of each individual in the population Repeat
Select best-ranking individuals to reproduce
Breed new generation through crossover and mutation (genetic operations) and give birth to offspring
Evaluate the individual fitnesses of the offspring
Replace worst ranked part of population with offspring Until <terminating condition>
```

Exhaustive *algorithm*

```
Exhaustive(intsol,intdepth)
{if(issolution(sol))
Printsoulution(sol)
Else
{solgenerated=generatesolution()
Exhaustive(solgenerated, depth+1)}
```

Covering algorithm

```
Inputs: labelled training dataset D
Outputs: rule set R that covers all instances in D
Procedure:
Initialize R as the empty set
For each class C{
While D is nonempty{
Construct one rule r that correctly classifies some instances in D that belong to class C and does not incorrectly classify any non-C instances
Add rule r to ruleset R
Remove from D all instances correctly classifiedby r}}
return R
```

LEM2 algorithm

```
Procedure LEM2
(input: a set B, output: a single local covering J of set B);
Begin
G:=B;
J :=$\emptyset$;
while G $\neq \emptyset$;
T :=$\emptyset$;
T(G) :={t|[t]$\cap G \neq \emptyset$};
while T = $\emptyset$ or [T] $\subseteq$ B
begin
select a pair t $\in$ T(G) with the smallest cardinality of [t];
if another tie occurs, select first pair;
$T \coloneqq T \cup \{t\}; G \coloneqq [t] \cap G;$
$T(G) \coloneqq \{t|[t] \cap G \neq \emptyset\};$
$T(G) \coloneqq T(G) - T;$
End {while}
for each t $\in$ T do
if $[T - \{t\}] \subseteq B$ then $T \coloneqq T - \{t\}$;
$\mathbf{J} \coloneqq \mathbf{J} \cup \{T\}$;
$G \coloneqq B - \cup_{T \in T} [T]$;
end {while};
for each T $\in$ J do
if $\cup_{S \in \mathbf{J}-\{T\}} [S] = B$ then $\mathbf{J} \coloneqq \mathbf{J} - \{T\}$;
end {procedure} .
```

Results

The results of Pima data set experiments using rough set approach are discussed here.

Table 8 : Rules through reduct for PIMA data set

Algorithm	**No. Of Reducts**	**Length of Reduct Min**	**Max**	**Mean**
Exhaustive	**32**	**3**	**5**	**3.8**
Genetic	**10**	**3**	**4**	**3.4**

Table 9 : Rule generation for PIMA data set

Algorithm	**No. Of Rules**	**Length of Rules**			**Accuracy (%)**	**Coverage**	**Filtered rules**	**Length of Rules**		
		Min	Max	Mean				Min	Max	Mean
Exhaustive	16364	3	5	3.8	71.8	0.152	20	3	5	3.6
Genetic	5110	3	4	3.4	78.16	0.1	10	3	3.6	3.1

Table 5.10 : Rule generation – Direct for PIMA data set

Algorithms	**Rules**	**Filtered Rules**	**Accuracy (%)**	**Coverage**	**Std. Dev.**
Exhaustive	5861	855	67.2	1	-
Covering	357	150	64.4	0.734	-
Lem2	300	114	76	0.293	-
Genetic	3574	749	64.26	0.990	3.14

The results of RSES implementing genetic algorithms to generate reducts are non-deterministic showing diverse accuracies for different executions for a single dataset. For these reasons the calculations are carried out multiple times and the average accuracy along with the standard deviation is considered and this procedure is followed from tables 8 to 10. The results of RSES implementing genetic algorithms aren't available anywhere.

Table 11 : Rule generation-Direct for PIMA data set (length of rules)

Algorithms	**Length of Rules**		
	Min.	**Max.**	**Mean**
Exhaustive	1	4	2.1
Covering	1	1	1
Lem2	2	6	3.5
Genetic	1	4	2

Table 12 : Optimal Cost prediction for PIMA data set

SI.No	Algorithm	Original Features	Features Reducted	Accuracy (%)	ROC (%)	Cost (units)	Average Cost (%)	Time (ms)
1	**With GA**	8	4	74.8	79.1	193	25%	0.02
2	**Without GA**	8	-	73.8	75.1	201	26%	0.06

Conclusion- Model 2

A multi-criteria optimization problem is introduced by feature subset selection and this in turn assists meaningful pattern recognition. This is strongly dependent on the attribute selection describing patterns. A subset of input variables is selected and those with neither minimal nor no predictive information are eliminated. Since its introduction in 1982 Rough set theory has proven to be an effective technique for data mining and knowledge discovery. Rough set algorithm deals with inconsistent data and generates decision rules helping the current study. The following conclusions are made:

(i) Exhaustive algorithm produces more number of reducts.

(ii) GA has more accuracy despite lesser coverage compared to Exhaustive algorithm.

(iii) The highest accuracy is that of LEM2 algorithm despite less coverage.

(iv) 4 number of features reduced, 74.8% accuracy, 0.02 ms is the time, 79.1% ROC value and 25% optimal cost with and without GA.

(v) Non deterministic results are obtained with RSES implemented GA. Therefore, accuracy is achieved by running the algorithm multiple times and getting the average of them. This is not found in literature and is first of its kind.

References:

1) Aboul Ella Hassanien and Jafar, M.H. Ali, "Rough Set Approach for Generation ofClassification Rules of Breast Cancer Data", Informatica, Vol. 15, No. 1, 23–38, 2004.

2) Aggarwal C.C. and Yu P., "Outlier detection for high dimensional data", Proceedings of ACM SIGMOD International Conference on Management of Data, 2001.

3) Ali K. M. and Pazzani M. J., "Error Reduction through Learning MultipleDescriptions", Journal of Machine Learning, 24: 3, 173-202, 1996.

4) Alvin C. Rencher,"Methods of Multivariate Analysis",WileyInterscience, second edition, 2002.

5) AmardeepKaur, Dr. Bhatia M.P.S., Dr. Bhaskar S.M., "State Of the Art of OutlierDetection in Streaming Data", IADIS European Conference Data Ming, 2007.

6) Cheick-OumarBagayoko,Jean-Charles Dufour, Saad Chaacho3, Omar Bouhaddou, Marius Fieschi, "Open source challenges for hospital information system(HIS) in developing countries: a pilot project in Mali", Bagayoko et al. BMC MedicalInformatics and Decision Making,10-22, 2010.

7) **Chien-Tsai Liu , Pei-Tun Yang , Yu-Ting Yeh, Bin-Long Wang,** "The impacts of smart cards on hospital information systems--an investigation of the first phase of the national health insurance smart card project in Taiwan", International Journal of Medical Infromatics,173-81,2006.

8) **Ching Wei Wang,** "New Ensemble Machine Learning Method for

Classification and prediction on Gene Expression Data", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, 2006.

9) **Chiu A. L and Fu A. W**., "Enhancement on local outlier detection", Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03), pg 298-307, 2003.

10) **Cover T.M and Hart P.E**., "Nearest neighbors pattern classification," IEEE Trans on Information Theory, vol. 13, pg 21-27, 1967.

11) **Cristina G. Dascălu, CorinaDimaCozma, Elena Carmen Cotrutz**, "Observations about the Principal Components Analysis and Data Clustering Techniques in the Study of Medical Data", World Academy of Science, Engineering and Technology 17,pg 69-73, 2006.