Feature Selection for Cancer Classification using Relative Decision Entropy

Ms. L. Meenachi, Dr. S. Ramakrishnan, R. Nivetha, R. Pavithra, S. Vasuki, S. Veena Priya Darsini

Department of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

Abstract**—***Features are the basis on which grouping is done to produce accurate results. The main goal of the feature selection is to determine the minimal features that are more efficient and could render high accuracy when compared with the whole set of features. Rough sets could only be considered as the effective tool for feature selection. Nowadays feature selection algorithms based on rough sets are prevailing. When they are considered on some analysis, it reveals that they could make high cost and much time to be worked out, as it suffers from intensive and exponential computation. For the purpose of eliminating the disadvantagesof these existing algorithms, a new algorithm called Feature Selection Models with Relative Decision Entropy is proposed. This algorithm is mainly based on roughness and degree of dependency, that includes both positive and boundary region calculation. It implies that this algorithm could provide good scalability for large data sets and lessens the cost along with the computation time.*

Keywords—Cancer Classification, Relative Decision Entropy, feature selection, Data mining

I. INTRODUCTION

Data mining is a wide area that deals with mining of data, where data would be presented in the form of features (attributes). Instead of increasing the features, some minimal features could be enough to produce finer results. The number of features doesn't matter, way of usage means a lot. So, we have to select features which could provide the best solution for the given problem. It deals with the identification of minimal features that are quite sufficient to produce highly reliable and most accurate results.

Rough set theory [3] is developed to meet the practical needs that can promote better classification and refined data. It finds application not only in data mining, it could be also used for the purpose of Intrusion Detection System to filter out the information that is redundant. It is used for the purpose of reducing large volumes of data. There are some limitations to propose a theory. According to rough set theory, it will change the underlying representation of the feature set. So that it could be quite readable and understandable for explaining it to the user. Many methods are proposed under rough set theory concepts. When there are some predefined methods, some new methods are designed to meet the needs that are not fulfilled by earlier methods. So, some heuristic methods even though can eliminate exponential powers, it would take a lot of time for computation purpose.

By algorithms, there are various forms of entropy: information and relative decision entropy. Information Entropy follows distinguishable discernibility matrix, but as time complexity increases, it is unendurable for handling large data sets. So, it is necessary to propose a new heuristic algorithm to handle required amount of datasets, provide accurate results with better classification performance and less time complexity.

An algorithm on Feature Selection with Relative Decision Entropy [7], involving computation of roughness and degree of dependency is handled for feature selection in the paper.

II. RELATED WORKS

Normally, there are two kinds of methods in rough set based feature selection theory. They are Hill climbing and Stochastic Methods .The hill climbing method involves the significance of the features. Based on the metrics, it is divided into three sections as positive region, discernibility matrix and information entropy based. Under the hill climbing methods we include, a new algorithm which is an extension of Shannon's information entropy, called the relative decision entropy with two basic concepts that is with *roughness and degree of dependency* [5]. Stochastic methods also have increased computational effort. So, many stochastic algorithms are proposed to lessen the computational complexity. Previously, an algorithm called FSRDE had been implemented. By currently, by enhancing certain necessary credentials, Feature selection with relative decision entropy is designed.

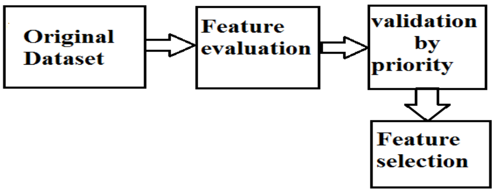

Fig.1: Block diagram of Feature Selection

2.1 Advantages of Feature Selection with Relative Decision Entropy

On the account of relative decision entropy, it deals only with the roughness concepts. Roughness is a metric to measure the uncertainty of the rough sets. Roughness can only handle information in the boundary region, but it does not consider regarding the positive region.

So, Inorder to eliminate the risks of relative decision entropy, we propose the new algorithm that deals with both degree of dependency and roughness. As roughness deals with the boundary region [8], similarly degree of dependency deals with the positive region. On combining both the results of boundary and positive region, more accurate results could be derived when compared to earlier algorithms.

In relative decision entropy method, with the given decision table DT= (U,C,V,D,f) for B contained in Union of C and D, we use a method on sorting to calculate the partition of Indiscernibility matrix U\IND(B). But in order to decrease the time complexity, we propose an incremental algorithm to compute U\IND (B).Thus the time complexity of Feature selection with relative decision entropy (O(|C U D|^2 * |U|)) is less than that of relative decision entropy
(O(|C U D|^3 * |U|))

Mainly, relative decision entropy [1] is not suitable for handling large data sets. It is sufficient only for small and low dimensional data sets. As information gets multiplied day by day, the size of the datasets increases. So, Feature selection with relative decision entropy could be the best algorithm for dealing with large data sets. This could provide good scalability and better accuracy even for multiple large datasets. Feature selection with relative decision entropy could provide high classification accuracies when compared to FSRDE. The relative decision entropy would decreases if the equivalence classes become smaller through finer partitioning methods. Also, feature selection with relative decision entropy could serve well for solving real world problems.

III. PROPOSED SYSTEM

Feature Selection is a tedious task, which involves computation of diverse characteristics of a data set. For efficient feature selection, several algorithms are to be imposed on the dataset. Computation of relative decision entropy involves intense computations, Information gain is one among it. Here, to apply information gain to a dataset, Info gain Attribute Evaluation algorithm [4] is used, which would evaluate the attributes in terms of its information gain value.

Information Gain approach could guarantee that test for attribute selection could be minimized and a simple decision tree could be formed. It ranks the attributes by priority with the use of ranker search algorithm that involves individual attribute evaluations. Later, properties of rough sets such as lower and higher approximation, roughness and degree of dependency are evaluated to minimize instances. The selected features are used for classification purpose. It can provide better accuracy than earlier methods.This could be achieved with the help of connecting Weka with Netbeans IDE. Weka acts as a Graphical User Interface which could produce resultant outputs of the database.

3.1 Calculating Information Gain

Computation of information gain is highly necessary. The attribute of highest information gain is used for partition of the subset. The attribute could represent the randomness or impurity in the partition. Here, the information gain could be calculated using the formula,

The expected information to classify the attributes in dataset is:

$$\textbf{Info (D)} = -\sum_{i=1}^{m} p_{i\,log_2(P_i)} \quad \textbf{[5] ... (1)}$$

After the portioning of the data set into subsets, to arrive at the exact classification, the equation would be:

$$\textbf{Info}\ _A(D) = \sum_{j=1}^{y} |Di|/|D| * Info(D_j)$$

$$\textbf{Gain (A) =Info (D) -Info}\ _A(D)\textbf{[5] … (2)}$$

3.2 Computation of Lowerand Upper Approximation:

For a given information system we have **IS=(U,A,V,f)**

Where,

U–an non empty finite set of objects

A-an non empty finite set of attributes

V-union of attribute domains

f-information function

For a decision table DT=(U,C,D,V,F) for any B $\underline{C}$ (C U D) and X $\underline{C}$ U, the lower andupper approximation of B in any subset X is:

$\underline{X}_B = U\{[X]_B{}^{\varepsilon}U/IND(B):[X]_B\underline{C}X\}$ **[5] …(3)**

$\underline{X_B} = U\{[X]_B{}^{\varepsilon}U/IND(B):[X]_B\frown X\neq Ø\}$**[5]…(4)**

3.3Roughnessof Approximation

For a decision table DT=(U,C,D,V,F) for any B $\underline{C}$ (C U D) and X $\underline{C}$ U(X! =0), the roughness $P_B(X)$ with upper and lower approximation are:

$P_B(X)=1-|\underline{X}_B|/|X_B|=|BND_B(X)|/|X_B|$**[5]… (5)**

The probability of roughness should be between **0 to 1.**The roughness would specify the amount of uncertainty of the roughest. A roughness of 1 would show that there is no certain knowledge of the set and 0 shows that everything is certain within the set.

3.4 Positive Region

$POS_B(D)=U_{X£U/IND(D)}X_B$**[5]... (6)**

D-set of all objects

U-non-empty finite set of objects

3.5 Degreeof Dependency

$Y_B(D)=|POS_B(D)|/|U|$**[5]… (7)**

Y_B (D)-ratio of all elements of U

if Y_B (D) =1 then D depends totally on B

if Y_B (D) <1 then D depends partially on B

3.6 Relative Decision Entropy

$RDE(D,B)=(1-Y_B(D))X\sum_{i=1}^{m} pB(D_i)\log_2 (p_B(D_i)+1)$**[5] …(8)**

Y_B (D)-degree of dependency of D and B

p_B ($D_{i)}$-roughness of set D_i

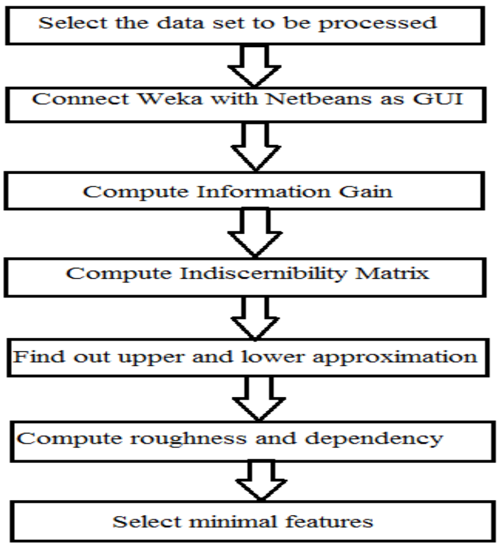
Select the data set to be processed
Connect Weka with Netbeans as GUI
Compute Information Gain
Compute Indiscernibility Matrix
Find out upper and lower approximation
Compute roughness and dependency
Select minimal features

Fig.2: block diagramof relative decision entropy algorithm

IV. RESULT

Initially a dataset is loaded into the weka tool. The dataset has a list of features which are to be minimized for effective results. Here, in the rough set, an algorithm is to be devised to produce better results than earlier algorithm. An algorithm called Relative Decision Entropy is proposed to provide large scalability for even large datasets. For implementing Relative Decision Entropy algorithm, computation of roughness and degree of dependency is necessary. The indiscernibility matrix must be calculated initially.

So, for this purpose, Information gain is to be found out for the partitioning of the Tuples, so that Relative decision entropy could be calculated based on the information gain.

To minimize the time complexity of the Relative Decision Entropy algorithm, an incremental algorithm to compute the partition of the Universe U is proposed. Experimental results demonstrated theeffectiveness of the Relative Decision Entropy in IDS and other application domains. Especially, Relative Decision Entropy performs better then FSRDE in terms of reducts, classification performance and running time.

Based on the algorithmic computation, the subset is generated. With a candidate set onthe subset returns selected features which could provide less computation cost, also eliminates the intensive computation. Thus efficient features for cancer classification could be found.

The attributes are reduced, with selection of minimal features on applying the Information gain.

Table 1: Feature selection using Infogain attribute evaluator

Types of cancer	Before feature selection	After feature selection
Breast cancer	10	9
Lung cancer	57	56
Leukemia	7130	7129

Table 2:Classification using decision table

Data set	Breast cancer	Lung cancer	Leukemia
Accuracy	96.875	99.856	97.0588
F measure	0.968	0.999	0.97
Sensitivity	1	0.9978	0.952
Specificity	0.958	0.9958	1
Kappa Statistics	0.92	0.9968	0.9386

V. CONCLUSION

The new Relative Decision Entropy algorithm that deals with both degree of dependency and roughness. As roughness deals with the boundary region, similarly degree of dependency deals with the positive region. On combining both the results of boundary and positive region, more accurate results could be derived when compared to earlier algorithms. Relative Decision Entropy could be the best algorithm for dealing with large data sets. This could provide good scalability and better accuracy even for multiple large datasets. RelativeDecision Entropy could provide high classification accuracies when compared to FSRDE. The relative decision entropy would decreases if the equivalence classes become smaller through finer partitioning methods. Also, relative decision entropy could serve well for solving real world problems. The time complexity of relative decision entropy (O (|C U D|^2 * |U|)) is less.

Thus the new algorithm on feature selection with relative entropy (Relative Decision Entropy) provides

Less computation cost

Better classification of features

Good scalability of features

High efficiency

Relative Decision Entropy was proposed based on Pawlak's classical rough set model and to deal with continuous features, the classical rough set model should replace all continuous features with discretized features by the process of discretization. However, discretization may cause the loss of information. Hence, in future work, we plan to extend Relative Decision Entropy to the neighborhood rough set model or the fuzzy rough set model .These extended rough set models can deal with continuous features without discretization.

REFERENCES

[1] Y.H.Qian, J.Y.Liang, W.Pedrycz, C.Y.Dang,"An efficient accelerator for attribute reduction from incomplete data in rough set framework",Pattern Recognition.44 (8) 1658–1670. (2011)

[2] [2] H. Yan, X.T.Yuan, S.C.Yan, J.Y.Yang, "Corr entropy based feature selection using binary projection", Pattern Recognition. 44(12) 2834–2842. (2011)

[3] Q.H.Hu, D.R.Yu, W.Pedrycz, D.G.Chen, "Kernelized fuzzy rough set sand their applications", IEEETrans.Knowl.DataEng.23

[4] H.S.Nguyen,S.H.Nguyen, "Discretization methods in data mining" ,in: L. Polkowski,A.Skowron(Eds.), ,Physica, Heidelberg, ,pp.451–482.(1998)

[5] R.Jensen, Q.Shen, "Semantics-preserving dimensional it reduction rough and fuzzy-rough based approaches", IEEE Trans.Knowl.DataEng.16(12) 1457–1471. (2004)

[6] P.Maji, S.K.Pal, "Feature selection using information measures in fuzzy approximation spaces", IEEE Trans. Knowl. Data Eng.22 (6)854–867. (2010)

[7] http://archive.ics.uci. edu/ml

[8] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html