Agris on-line Papers in Economics and Informatics

Volume VI Number 4, 2014

Rainfall Forecast Analysis using Rough Set Attribute Reduction and Data Mining Methods

M. Sudha, B. Valarmathi

School of Information Technology and Engineering, Vellore Institute of Technology - University, India

Abstract

Developments in information technology has enabled accumulation of large databases and most of the environmental, agricultural and medical databases consist of large quantity of real time observatory datasets of high dimension space. The curse to these high dimensional datasets is the spatial and computational requirements, which leads to ever growing necessity of attribute reduction techniques. Attribute reduction is a process of reducing the data space by removing the irrelevant, redundant attributes from large databases. The proposed model estimates the enhancement achieved in spatial reduction and classifier accuracy using Rough Set Attribute Reduction Technique (RSART) and data mining methods. The first module of this proposed model has identified an efficient attribute reduction approach based on rough sets for spatial reduction. The next module of the proposed model has trained and tested the performance of Naive Bayes (NB), Bayesian Logistic Regression (BLR), Multi Layer Perceptron (MLP), Classification and Regression Tree (CART) and J48 classifiers and evaluated the accuracy in terms of each classification models. The experimental results revealed that, the combination of RSART based on Genetic Algorithm approach and Bayesian Logistics Regression Classifier can be used for weather forecast analysis.

Key words

Attribute reduction, rough set, genetic algorithm, Bayesian Logistics Regression, rainfall forecasts, classification, accuracy.

Introduction

Rough set theory was proposed by Zdzisław Pawlak in 1982 as a mathematical tool for data analysis; it is a different mathematical approach to handle vagueness and imperfect knowledge. Pawlak (1982) stated that Indiscernible relation is the mathematical basis of rough set theory, i.e. every object in the universal set has some information, objects characterized by same values are indiscernible in view of the available information and objects with different values are discernible. Any non-empty set of all indiscernible objects is called an elementary set and forms a basic grain of knowledge about the universe. Pawlak (1982) described that the union of some elementary sets is defined as a crisp set otherwise the set is rough. Each rough set has boundary region with objects which cannot be classified to a particular set. In any rough set a pair of precise sets, called the lower and the upper approximation is the grain of rough set. The lower approximation consists of all objects which certain to be a member of a set and upper approximation has the possible members of the set. The upper and the lower approximation difference constitute the boundary region elements in a rough set (Grabowski, Jastrzebska, 2009).

Rough set based on data analysis starts from a data table called a decision table, columns of which are labelled by attributes, rows or tuples by objects of the table are attribute values. Attributes of the decision table consists of disjoint groups called condition and decision attributes. Rough sets have been used in various medical, meteorological applications for knowledge discovery (Shen, Jensen, 2007). In this approach all manipulations are performed on the corresponding data itself without any prior perception of any additional information about data description.

Some more essential concepts in rough sets theory are the cores and reducts. For instance let {S} be a subset of universal set, {X} is an information system let {A} be the attribute superset that containsthe complete set of conditional attributes and decision attribute now the reduct is defined as {R} (Pawlak, Skowron, 2007) {R} be the subset of attribute superset {A} that has the most significant attributes, then the R-lower approximation { S} is the set of all elements of {S} which can be with certainly

classified as elements of a specified concept. Let R-upper approximation set of S is a subset of {X} and R is non null set with objects that are possible members of specified concept. Approximation space of {S} is determined by information contained in {B} using B-lower approximation space and B-upper approximation space w.r.t to {S}. Approximation set are represented as $\underline{B}S$ for lower and $\overline{B}S$ for upper.

Lower approximation set of S, is the set of elements of X which can be with certainty classified as elements of S w.r.t a specified concept, R-lower approximation set of S is defined as $\underline{B}S = \{x \mid [x]_B \subseteq S\}$, upper approximation set of S, is set of elements of X which can possibly be classified as belonging to the set S w.r.t a particular concept R-Upper approximation set of S is defined as $\overline{B}S = \{x \mid [x]_B \cap S \neq \phi\}$.

Boundary region represents the uncertain portion of the dataset, with less information about the datasets to definitely establish the class of data (Pawlak, 2002). Boundary region is defined as [{upper approximation} – {lower approximation}]. A reduct set contain significant set of attributes of superset {A}, attributes in a reduct set {R} are more predictive of a given decision variable as that of {A}. Reduct set are generated based on indiscernibility matrix, discernibility matrix, boundary region elements, positive region elements and the equivalence partitions of rough set theory.

Lower and upper approximation space is determined using the cardinality of approximation space, the accuracy is defines as below

$$\alpha_B(S) = \frac{|\underline{B}(S)|}{|\overline{B}(S)|}$$

For {S}, if the accuracy α(s) = 1 then S is a crisp set with respect to B. If α(s) < 1 , then S is a rough set with respect to B with imperfect knowledge. Core in rough set theory consists of significant attribute that it is most predictive feature of a decision table, core is determined as indispensable attribute of {A} belonging to {R}. Identification of coreattributes is a significant task in knowledge processing. Processing raw data directly is not advisable as it may affect the quality of the data analytics. In this situation, the process of attribute reduction that identifies the significant attribute of {A} play a key role by removing the redundant, irrelevant attributes. (Wei et al., 2012) described the application of rough set concept for hybrid data which involve different data with imperfect knowledge can be handled efficiently using rough set. (Greco et al., 2001) described about the use of rough sets concept for multi criteria data analysis. The reduct {R} generated by attribute reduction algorithm is then subject to classification. Classification is a two step process consisting of training phase and testing phase. It constructs a classifier model by learning from a training set. (Yu et al., 2005) described a new classification approach by integrating feature selection algorithms to enhance predictor accuracy.

In testing phase, the model is tested with unseen samples and the classifier's accuracy is determined. If the test samples are the disjoint records that are randomly selected from the data set independent of the training samples then accuracy of a classifier for a given test set is the percentage of test set samples that are correctly classified. The associated class label of each test record is compared with the learned classifiers class prediction for that record. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data records for which the class label is not known.

In our research investigation we have used Rough Set for attribute reduction to generate the reduct sets based on Johnson algorithm and Genetic Algorithm approach. From our results we realized that, the irrelevant attributes can pull down the prediction accuracy and can even mislead the learning process.

Literature review

(Yao, 2009) stated that rough sets discernibility matrix based on attribute reduction approach has been adopted widely to find reduct sets for different applications and they have demonstrated the importance of attribute reduction using some sample dataset. (Li et al., 2010) described that elementary matrix simplification operations and introduced operations to transform matrix into a simpler representations. Have emphasized that the elements of positive region are of great importance and it is necessary to verify that the objects in positive region are never missed in attribute assessment.

(Qablan et al., 2012) shown a new attribute reduction approach based on Ant Colony Optimization algorithm. The application of this method confirms that the computation is reduced and the results are good using this algorithm compared with the traditional algorithm. Thus, it is proven that this method is a fast and efficient algorithm

of attribute reduction. Modified binary discernibility matrix and attribution reduction algorithm based on binary discernibility matrix is an ordering approach with a new simple link concept in the algorithm has supported to reduce the size of the table to reduce the computation and storage complexity.

A reduct optimization method based on the condition attributes has been discussed; this can classify the grouping generated representative data to simplify the discernibility matrix, and the order of the discernibility matrix, and the complexity of the attribute reduction (Miao et al., 2009). Johnson reduction algorithm and the Object Reduct using Attribute Weighting technique algorithm (ORAW) for reduct computation are some simple widely applied reduct and rule generation techniques. (Suguna et al., 2011) have described presented a new feature selection method based on rough set approach integrated with the Bee Colony Optimization (BCO). This proposed approach generates minimal reducts for medical datasets.

(Sudha, Valarmathi, 2013) and (Suguna et al., 2011) have mentioned that attribute reduction approach based on quick reduct, entropy measure based on reduct, hybrid rough set based on genetic algorithm), Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) were widely adopted for attribute reduction. Experimental results conclude that rough set bases techniques achieved optimal reduction and have generated better number of reducts than other methods. Rough Set attribute assessment based on Genetic Algorithm approach is recommended for the numerical datasets in order to achieve global optimization.

Analogical matrix based on attributed reduction algorithm, a totally new approach towards attribute reduction has been discussed from experimental evaluation reveal that it can reduce time complexity and spatial complexity without breaking the coherence of information contained in decision table (Mahapatra et al., 2010).

(Bal, 2013) stated that a two-step attributes reduction technique based on Bayesian model for decision theoretic rough set using different classification properties. The first step creates reconstruction of the complete decision table; the second step determines the optimal reducts for data analysis. A new Scan vector approach based on attribute reduction; the proposed method defines a new conception of discernible vector by which the discernible matrix information table can be transformed into discernible vector set (Xu E et al., 2006).

(Wei et al., 2012) introduced Attribute Reduction of Decision Table based on attributes similarity relation. Hao and Zhang (2010) have described about reduct optimization method based on the condition attributes that classify the grouping generated representative data to simplify the discernibility matrix, and the order of the discernibility matrix, and the complexity of the attribute reduction. Thus the time complexity and space complexity made optimization, save the time and space complexity.

(Huang et al., 2013) introduced a new algorithm of attribute reduction using the analogical matrix, and the correctness and feasibility of it was proved. The algorithm can reduce time complexity and spatial complexity of attribute reduction, and do not break the coherence of information contained in decision table. The analysis of the realistic example shows that the algorithm is effective and feasible.

(Olaiya et.al , 2012) mentioned that the following approaches based on Artificial Neural Network, decision tree, Genetic Algorithms, Rule Induction, Nearest Neighbor method, memory-based reasoning, logistic regression and discriminant analysis are widely adopted for predictive data mining tasks. Taking advantage of these models they show that the Artificial Neural Network approach and decision tree were used for rainfall forecast analysis to study the climate change.

Valmik and Nikam et al. (2013) have described a rainfall prediction model based on Bayesian classifier. In Bayesian approach perform well for those datasets with predictor class label however in the absence of predictor class label for a given dataset the bayesian classification model assumes the record with zero probability there by affecting the overall accuracy.

Materials and methods

This observatory meteorological data has eight parameters: Max (maximum temperature), Min (minimum temperature), RH1 (relative rumidity-1), RH2 (relative humidity-2), Wind, SR (solar radiation), SS (sun shine), and EVP (evapotranspiration). The above mentioned parameter‘s values are recorded in an observatory in daily basis as per the standard norms. For our experimental analysis we have used around 10,000

days recorded observatory records. Materials used in recording the atmospheric values were thermometer for temperature measurement, robinson cup anemometer for recording the wind speed, pirenometer for recording solar radiation, sun shine recorder strip to record sunshine andevaporimeter for recording evapotranspiration. In our research work we have used a real time observatory post rainfall predictions weather data.

The targeted sample input is in the form of the Table1 as defined in rough set theory. The parameters other than RF (Rain Fall) are referred as condition attribute and RF is the decision attribute w.r.t Rough Set theory approach, the reduct set {R} contains the significant attributes of set {A}, where {A} is the complete attribute set of a given input table. Set {R} is a reduct, which is our input for the classification module. The sample format of our dataset is represented in Table 1. In this research, our initial focus is to identify the most significant attributes from the complete attribute set {A} which results in identification of essential attribute subset {R}. The identified minimal attribute sets in turn will reduce the data space, computational complexity and it can improve the prediction accuracy.

Rough Set Attribute Reduction Techniques

The proposed model incorporates rough set based on attribute reduction techniques. In general identification of significant attribute is one the significant task knowledge representation. The input for attribute reduction module is the set of tuples of the data table and corresponding attribute set inclusive of conditional and decision attribute (class label).

Rough Set based Johnson Algorithm (JA)

Johnson algorithm is widely adopted for finding the significant attribute set from the complete attribute list; the algorithm has been evaluated and implemented using Java and Rosetta software. The Johnsons reduct starts with initialization of an empty Reduct Set R-{} followed by step 1-6

1. For each Row r_i in Discernibility Matrix compute attribute with Maximum Frequency, add the attribute to the Reduct set R{}
2. While there are still entries left in r_i.
3. Add the attribute (a) with maximum frequency to Reduct Set.
4. If more attributes have the same maximum frequency then chose any one at random
5. Delete all entries that contain attribute (a) from r_i.
6. End.

Rough Set based Genetic Algorithm (GA)

Rough Sets based Genetic Algorithm, the solutions can be coded as strings of 0's and 1's. An initial population of solutions is generated randomly and the best solutions, according to some fitness function, are iteratively chosen to breed new generations of solutions using genetic operators such as mutation and crossover. Algorithm encodes potential solution candidates referred as chromosomes and the set of potential solution candidates labelled as generation. Actual realization starts with a population of chromosomes and set of algorithm based on operators determines

MAX Celsius	MIN Celsius	RH1 %	RH2 %	WIND km/hrs	SR KCalories	SS hrs	EVP mm	RF Class Label
28.0	16.0	95.0	42.0	8.6	243.2	10.2	4.4	0
28.5	16.5	85.0	41.0	9.0	241.6	10.4	2.8	0
28.5	17.5	95.0	52.0	8.0	183.2	7.0	5.0	0
28.0	17.5	92.0	57.0	6.1	209.6	8.6	2.6	0
28.0	14.0	94.0	55.0	7.4	260.0	10.2	3.4	0
28.0	18.0	95.0	51.0	9.0	232.0	9.4	4.2	0
35.0	23.6	85.0	43.0	7.0	214.4	7.6	5.1	1
33.0	23.0	90.0	69.0	6.4	332.3	6.2	3.9	1
29.0	22.0	91.0	48.0	3.5	477.0	8.7	3.4	1
31.5	22.5	90.0	56.0	10.6	432.1	7.8	4.4	1
31.0	22.0	90.0	58.0	5.1	352.5	7.6	2.6	1

Source: TNAU Coimbatore India

Table 1: Observatory record of rainfall prediction (1984 – 2012). Dataset.

the better solutions. Genetic Algorithm is a population based model that makes use of selection and recombination operators to find better solutions in the search space. In this Rosetta software this genetic algorithm approach is implemented as supervised learning, it involves a number of search problems that may easily be approached with heuristic search.

Proposed model

The assessment of suitable model for rainfall forecast analysis for optimal prediction accuracy consists of several processing stages. For clarity we have modularized the stages as two specific modules, the first module consists of five stages so as to generate the possible feature subsets using rough set approach hence module I is referred as attribute reduction module and next is the classifier identification module in which a suitable classifier is identified. Reduct sets are the input for this classifier identification module, it has two main sub stages model training and model testing phase followed by accuracy analysis of each classifiers.

Attribute Reduction Module

Step 1 - Data pre-processing.

Step 2 - Removal of outliers from raw data.

Step 3 - Apply RART on target data using Johnson algorithm approach.

Step 4 - Apply RART on target data using Genetic algorithm approach.

Step 5 - Generation of reduct sets {R}.

The reduct sets of Johnson algorithm and Genetic Algorithm were later evaluated by classifiers to analyze the enhancement in accuracy in the next module.

Classifier Identification Module

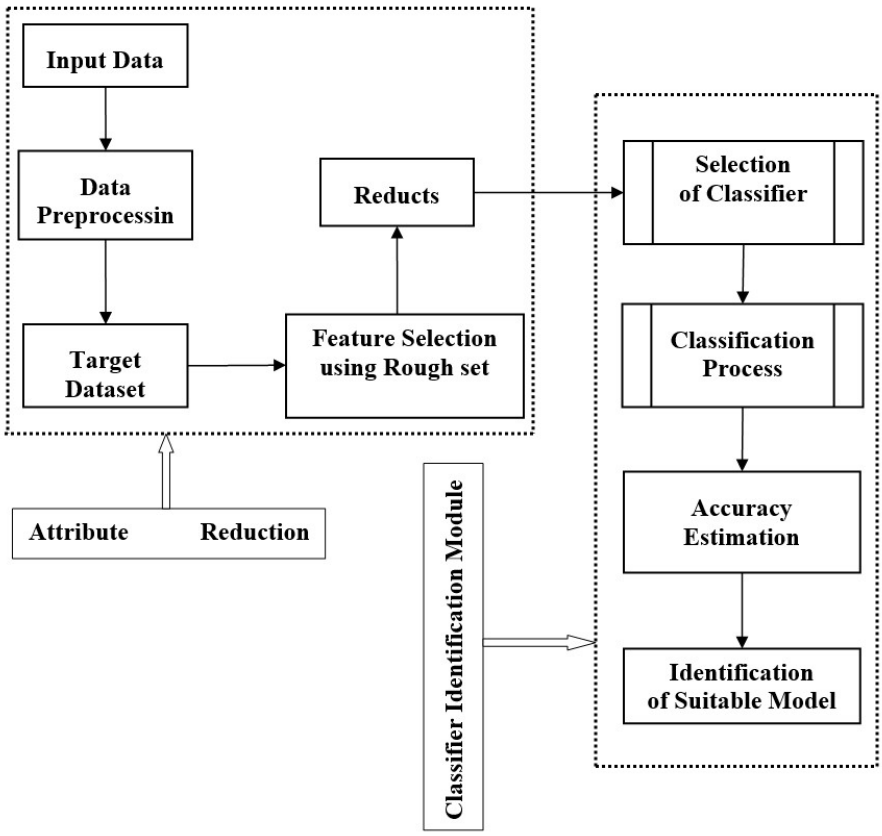
Step 1 - Input - reduct sets {R}.

Step 2 - Identify the Classifier.

Step 3 - Construct confusion matrix for each reduct.

Step 4 - Estimate the accuracy obtained.

Step 5 - Terminate the process.



Source: own processing

Figure 1: Rough Set and data mining based Rainfall Forecast Model

In classifier identification module we have analysed the performance of Naïve Bayes, Bayesian Logistics Regression, Multi-Layer Perceptron, Classification and Regression Tree (CART) and J48 classifiers to identify the suitable classification model that outperform for this meteorological dataset. The classifiers are trained and tested with the complete reduct sets {R1} to {R10} independently. Later all the reduct set accuracy is estimated using true positive and true negative classification. Evaluation process of the classifier is done using Java based weka3.11 software

Working Model of Classifiers

Naive Bayes Classifier, Bayesian Logistics Regression, Multi Layer Perceptron J48 and CART classification algorithms performance were estimated. These classification algorithms were used to analyze the real meteorological data registered between 1984 and 2013 post rainfall around the Coimbatore region of India. We have used Weka (Waikato Environment for Knowledge Analysis) software; a widely adopted suite of machine learning written in Java, developed at the University of Waikato for our classifier performance assessment. WEKA is free software available under the GNU General Public License. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality (Accessed: 23 Feb 2014). WEKA is used for analysing the selected classifiers and evaluation of module II were based on the testing set. A simple 10 fold cross validation was performed, as described in (Cao et al., 2013). So we have used 65% for training and the remaining 35% tuples for testing. The process was repeated ten times and the accuracy for true, false, and total accuracy estimated based on confusion matrix. The final accuracy is the average of the accuracy in all tests.

Naive Bayes Classifier

Bayesian Classifier is classifier based on conditional probability and the effect of an attribute value on a given decision is independent of the values of the other attributes. This hypothesis is called class conditional independence. Naïve Bayes Classifier predicts class membership probabilities such as the probability that a given tuples belongs to a particular class. Estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations in the training phase, the probability of each class is computed by counting how many times it occurs in the training dataset. Naïve Bayes Algorithm computes the probability for the instance X, given C with the assumption that the attributes are independent (Ramana, et al., 2011). Later the probabilities are calculated from the frequencies of the instances in the training set. During training, the probability of each class is computed by counting how many times it occurs in the training dataset. This is called the "prior probability" P(C=c). In addition to the prior probability, the algorithm also computes the probability for the instance x given c with the assumption that the attributes are independent, the algorithm attempts to estimate the conditional probabilities of classes given an observation as rainfall occurred or else rainfall has not occurred (Krishnaiah et al., 2013).

Bayesian Logistics Regression

Logistic regression is a discriminative linear classification model with some decision boundary. It has been proven that logistic regression shows higher accuracy when training data is large. Naive Bayes Classifier has shown good result when the training data size is small. In our process of classification we have used binary class label as RF = 1 or RF = 0, this RF is characterized by set of eight conditional attributes referred as Feature vector X, Bayesian approach with logistic regression is used in different field of application (McNevin et al., 2013), it has shown significant performance in medical data diagnosis in disease identification and it performs well for larger datasets (Miao et al., 2009).

Multi-Layer Perceptron Classifier

In a classification process, the outcome of MLP classifier is class membership for the given input reduct set. The advantage of a neural network algorithm is it adjusts themselves to the application by means of the training or learning process. MLP network-based classifiers have shown good results in application (Zhang et al., 2000). Multi-Layer Perceptron is feed forward neural network which is widely using in classification of data. In MLP the raining processing with the set of input values X and its target T makes use of an objective function such as error, cost or some function. O (Y, T), to find the deviation of the predicted output class labels, Z = MLP (Y; W) from the observed data value T and makes use of the assessed outcome to converge to an optimal set of weights W is based on the algorithm.

J48-Induction Tree

J48 is a classification algorithm based on induction tree algorithm; it uses information entropy and information gain measure for the splitting criterion. The attribute with the highest normalized information gain is chosen to make the decision and then recurs on subset. It constructs a decision tree starting from a training set T, training set is a set of rows in the rainfall dataset. The class label has only discrete values 0 means no rainfall and 1 means rainfall occurred. We denote with numeric values 0, 1 for the Class values of the class. The J48 algorithm constructs the decision tree with a divide and conquers strategy. According to J48 in each node in a tree is associated with a set of cases that are assigned with weights to take into account unknown attribute values. At the beginning, only the root is present, with associated the whole training set S and with all case weights equal to 1:0. At each node the divide and conquer strategy is applied in order to find the locally best option and backtracking is not allowed.

Classification and Regression Tree

Decision tree classification models are usually used in data mining to observe the data to induce the tree and its rules will be used to make predictions. In a decision tree each branch node represents an option between a number of alternatives, and each leaf node represents a decision. CART is widely used in determination and classification of medical diagnostics datasets (Nishida, Nobuko, 2005) and (Deconinck et al., 2005). Classification and Regression Tree (CART) generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed this is called a multiway tree (Data mining Models and Algorithms, 2014). This tree structure will enable to make easy and efficient class prediction.

Results and discussions

Rough sets based attribute reduction is carried out using Johnson's Algorithm and Genetic Algorithm approach. In Johnson approach only one reduct is generated which is not of scope. Rough Set based Genetic Algorithm has shown significant result for our meteorological data set. Table 2 projects the reduct sets with the significant attributes and the corresponding spatial reduction achieved using genetic algorithm approach.

Result analysis of Module-I

Reducts set	Significant attributes	Number of attribute in each reduct	Reduction achieved
{R1}	R1_12346	5 attributes out of * 8	37.5 %
{R2}	R2_12356	5 attributes out of * 8	37.5 %
{R3}	R3_12457	5 attributes out of * 8	37.5 %
{R4}	R4_13457	5 attributes out of * 8	37.5 %
{R5}	R5_13467	5 attributes out of * 8	37.5 %
{R6}	R6_23456	5 attributes out of * 8	37.5 %
{R7}	R7_23457	5 attributes out of * 8	37.5 %
{R8}	R8_23467	5 attributes out of * 8	37.5 %
{R9}	R9_2567	4 attributes out of * 8	50 %
{R10}	R10_4567	4 attributes out of * 8	50 %

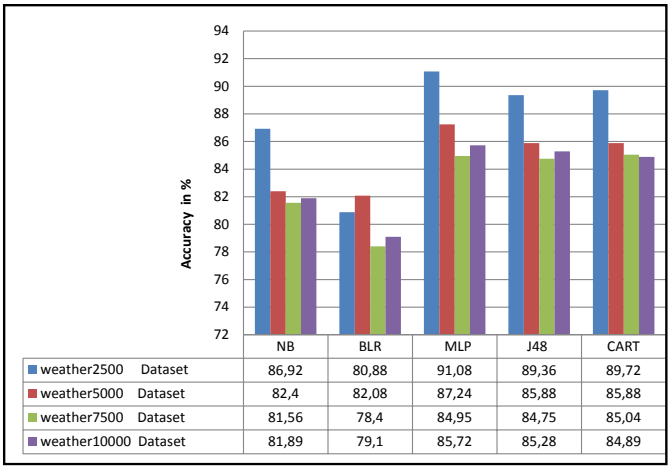
Source: own survey

Table 2: Reduct set attribute reduction

The attribute reduction achieved is shown in Table 2. The genetic algorithm has generated ten significant reduct sets {R1} to {R10} with 4 to 5 significant attributes out of 8. Spatial reduction achieved for every feature subset (reduct) is given in Table 2.Using this attribute reduction we have achieved significant data reduction of 37.5% and maximum of about 50 %of spatial.

Result analysis of Module II

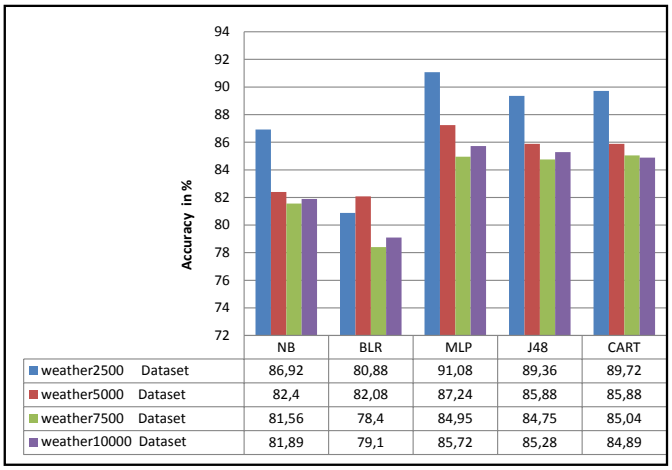
The Chart 1 and the Chart 2 projects the estimation of classifier accuracy before and after using rough set attribute reduction based on Johnson algorithm approach. The accuracy estimation [(Tp+Tn)/(Tp+Tn+Fp+Fn)]*100] of Genetic algorithm approach based on the confusion matrix is given in the Chart 3, all the reducts {R1} to {R10} is evaluated using all the five selected classifiers. Bayesian Logistics Regression (BLR) model has shown significant enhancement than other classifiers in terms of accuracy. BLR has achieved an improved accuracy on the following reduct sets {R1} = 80.19%, {R2} = 80.11%, {R3} = 80.43%, {R4} = 80.59% and {R7} = 79.95%. The experimental results of this proposed model has shown that Rough Set Attribute Reduction technique based Genetic algorithm (RSAT-GA) and Bayesian Logistics Regression (BLR) model has shown significant enhancement than other classifier.



	NB	BLR	MLP	J48	CART
weather2500 Dataset	86,92	80,88	91,08	89,36	89,72
weather5000 Dataset	82,4	82,08	87,24	85,88	85,88
weather7500 Dataset	81,56	78,4	84,95	84,75	85,04
weather10000 Dataset	81,89	79,1	85,72	85,28	84,89

Source: own processing

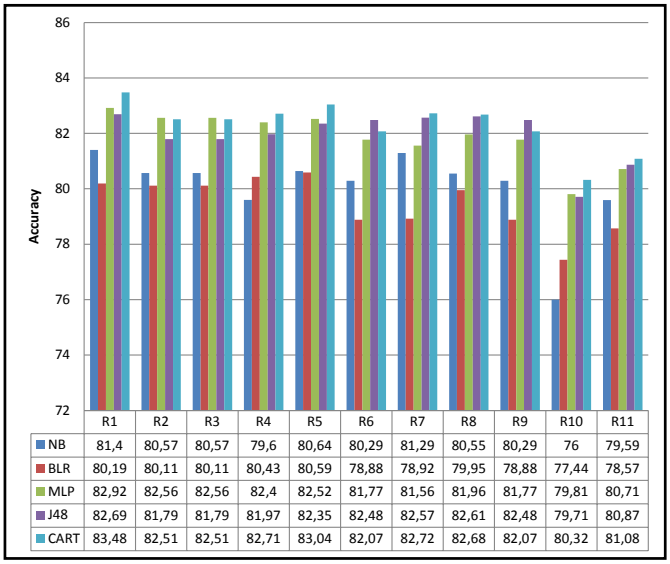
Chart 1: Accuracy before Attribute Reduction.



	NB	BLR	MLP	J48	CART
weather2500 Dataset	86,92	80,88	91,08	89,36	89,72
weather5000 Dataset	82,4	82,08	87,24	85,88	85,88
weather7500 Dataset	81,56	78,4	84,95	84,75	85,04
weather10000 Dataset	81,89	79,1	85,72	85,28	84,89

Source: own processing

Chart 2: Accuracy after Attribute Reduction using Johnsons Algorithm.



	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
NB	81,4	80,57	80,57	79,6	80,64	80,29	81,29	80,55	80,29	76	79,59
BLR	80,19	80,11	80,11	80,43	80,59	78,88	78,92	79,95	78,88	77,44	78,57
MLP	82,92	82,56	82,56	82,4	82,52	81,77	81,56	81,96	81,77	79,81	80,71
J48	82,69	81,79	81,79	81,97	82,35	82,48	82,57	82,61	82,48	79,71	80,87
CART	83,48	82,51	82,51	82,71	83,04	82,07	82,72	82,68	82,07	80,32	81,08

Source: own processing

Chart 3: Accuracy after Attribute Reduction using Genetic Algorithm

Conclusion

The main findings of our proposed model are listed as follows. (1) Rough Set Attribute Reduction Technique based Genetic Algorithm approach (RSAT-GA) has achieved optimal reducts for realtime meteorological (rainfall prediction) dataset with eight atmospheric parameters. (2) Bayesian Logistics Regression (BLR) have shown improved prediction accuracy than other classifier after attribute reduction. (3) This model is cost effective, simple and reliable. (4) It is suitable for larger datasets.

Future research directions

In future, we would like to evaluate our model with latest classification techniques and have proposed to incorporate latest dimensional reduction approaches like Map Reduce Paradigm along with data mining methods to achieve optimal enhancement in weather forecasts.

Corresponding author:
Prof.M.Sudha
School of Information Technology and Engineering
Vellore Institute of Technology - University, Vellore, India
Phone: +91 9443744781, E-mail: msudha@vit.ac.in

References

[1] Bal, M. Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision Table, Information Science Letters an Int. Journal, 2013, p. 35-47. ISSN 2325-0399.

[2] Cao, L., Liu, X., Wang, Z. P., Zhang, Z. The spatial outlier mining algorithm based on the KNN graph. Journal of Software, 2013, Vol. 8, 12, , p. 3158-3165. ISSN 1796-217X.

[3] Data mining Models and Algorithms, [Online] Avalibale: http://www.huaat.com/English/datamining/D_App.htm [Assessed: 2014-04-18].

[4] Deconinck, E., et al. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. Journal of Pharmaceutical and Biomedical Analysis, 2005, p. 91-103. ISSN: 0731-7085.

[5] Grabowski, A., Jastrzebska, M. Two Formal Approaches to Rough Sets, Studies in Logic, Grammar and Rhetoric, 2009, Vol. 18, 31, p. 25-34. ISSN:0860-150X.

[6] Greco, S., Matarazzo, B., Słowin´ski, R. Rough set theory for Multicriteria Decision Analysis, European Journal of Operational Research, 2001, p. 1–47. ISSN 0305-0483

[7] Hao, W., Zhang, X. A simplified Discernibility matrix of the attribute reduction method, International conference on Information Management, 2010. ISSN 1793-6411.

[8] Huang, Y., Chen, S. An Algorithm of Attribute Reduction Based on Rough Sets, Physics Procedia, 2013, p. 2025 – 2029.

[9] Krishnaiah, V., Narsimha, G., Subhash, N. Chandra, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, 2013, Vol. 4, 1, p. 39 - 45. ISSN 0975-9646.

[10] Li, J., Fan, X., Wang, X. An Improved Attribute Reduction Algorithm Based on Importance of Attribute Value, Elseveir procedia Engineering, 2010, p. 56-359. ISSN 1877-0428.

[11] Mahapatra, S. Sreekumar, Mahapatra, S. S. Attribute selection in marketing: A rough set approach, IIMB Management Review, 2010, p.16-24. ISSN 0970-3896.

[12] McNevin, D., Santos, C., Gómez-Tato, A., Álvarez-Dios, J., de Cal, M.C., Daniel, R., Phillips, C., Lareu, M. V. An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals, Forensic Science International: Genetics Supplement Series, 2013, p. 63–64. ISSN 1875-1768.

[13] Miao, D. Q., Zhao, Y., Yao, Y. Y., Li, H. X., Xu, F. F. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model, Information Sciences 179, 2009, p. 4140–4150. ISSN 1066-2243.

[14] Nikam, V. B., Meshram, B. B., Modeling Rainfall Prediction Using Data Mining Method A Bayesian Approach, Fifth International Conference on Computational Intelligence, Modelling and Simulation, 2013.

[15] Nishida, N., Tanaka, M., Hayashi, N., Nagata, H., Takeshita, T., Nakayama, K., Morimoto, K., Shizukuishi, S., Determination of smoking and obesity as periodontitis risks using the classification and regression tree method, Journal of Periodontology, 2005, p. 923-928. ISSN 00223492.

[16] Olaiya, F., Adeyemo, A. B. Application of Data Mining Techniques in Weather Prediction and Climate Change Studies, I. J. Information Engineering and Electronic Business, 2012, 1, p.51-59. ISSN 2074-9031.

[17] Pawlak, Z. Rough Sets and its Applications, Journal of Telecommunications and Information Technology, 2002, p. 7-10.

[18] Pawlak, Z. Rough sets, International Journal of Computer and Information Sciences, 1982, p. 341-356. ISSN 2074-9058.

[19] Pawlak, Z., Skowron, A. Rough sets: Some extensions, Information Sciences, 2007, p. 28–40. ISSN 0020-0255.

[20] Qablan, T., Al-Radaideh, Q. A., Shuqeir, S. A. A Reduct Computation Approach Based on Ant Colony Optimization, Basic Sci. and Eng., 2012, Vol. 21, 1, p. 29-40.

[21] Ramana, B. V., Surendra, M., Babu, P., Venkateswarlu, N. B. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems (IJDMS), Vol.3, 2, May 2011, p.101-114, ISSN: 0975-5705.

[22] Shen, Q., Jensen, R. Rough Sets, their Extensions and Applications, International Journal of Automation and Computing, January, 2007, p. 100-106. ISBN: 978-3-902613.

[23] Sudha, M., Valarmathi, B. Exploration on Rough Set based Feature Selection, International Journal of Applied Engineering Research, 2013, p. 1555-1556, ISSN: 1087-1090.

[24] Suguna, N., Thanushkodi, K. G. An Independent Rough Set Approach Hybrid with Artificial Bee Colony Algorithm for Dimensionality Reduction, American Journal of Applied Sciences, 2011, p. 261-266. ISSN 15543641.

[25] Wang, C.-Z., Cui, X.-H., Bao, W.-Y., He, Q. Attribute Reduction of Decision Table based on Similar Relation, International Conference on Machine Learning And Cybernetics, Xian, 2012, p. 15-17, ISSN: 2160-133X.

[26] Wei, W., Liang, J., Qian, Y. A comparative study of rough sets for hybrid data, Information Sciences, 2012, p. 1–16, ISSN 1793-641.

[27] Weka Software, http://www.cs.waikato.ac.nz/ml/weka/, [Assessed: 2014-02-23].

[28] Xu, E., Gao, X. D., Tan, W. D., Attributes Reduction Based On Rough Set, 5th International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August, 2006.

[29] Yao, Y. Discernibility matrix simplification for constructing attribute reducts, Information Sciences, Elseveir, 2009, p. 867–882, ISSN 0020-0255.

[30] Yu, L. Toward Integrating Feature Selection Algorithms for Classification Clustering, IEEE Transactions on Knowledge and Data Engineering, 2005, Vol. 17, 4. ISSN 1041-4347.

[31] Zhang, G. P. Neural networks for classification: a survey, IEEE Transactions on Systems, Man and Cybernetics, Part C 30, 2000, p. 451–462. ISSN 1094-6977.