Available online at www.elixirpublishers.com (Elixir International Journal)

Applied Mathematics

Elixir Appl. Math. 94 (2016) 40397-40399

ISSN: 2229-712X

Rough Set: Buzzword of Data Classification

Niharika Upadhyay[1,*] and Pragati Jain[2]

[1,*]Department of Science, Pacific University, India.

[2] Department of Science, St. Paul Institute of Professional Studies, India.

ARTICLE INFO

Article history:

Received: 5 March 2016;

Received in revised form:

9 May 2016;

Accepted: 13 May 2016;

Keywords

Rough Set Theory,

Naïve Bayes,

Random Forest,

k Star,

Multilayer Precepron and

j48.

ABSTRACT

Classification is an important Data Mining Technique with broad applications in every walk of life. It is termed as classifying each item in a set of data into one of predefined set of classes or groups. The present study compares the performance evaluation of Naïve Bayes, Random Forest, k Star, Multilayer Preceptron, j48 classification algorithm and Rough Set Theory. The paper presents the experimental results about classification accuracy and explores that the accuracy of Rough Set Theory is improved than other algorithms.

Introduction

The fundamental algorithms in data mining and analysis form the basis for the emerging field of data science, which includes automated methods to analyze patterns and models for all kinds of data, with applications ranging from scientific discovery to business intelligence and analytics.

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. Data mining comprises the core algorithms that enable one to gain fundamental insights and knowledge from massive data. It is an interdisciplinary field merging concepts from allied areas such as database systems, statistics, machine learning and pattern recognition. In fact, data mining is part of a larger knowledge discovery process, which includes pre-processing tasks such as data extraction, data cleaning, data fusion, data reduction and feature construction, as well as post-processing steps such as pattern and model interpretation, hypothesis confirmation and generation, and so on. This knowledge discovery and data mining process tend to be highly iterative and interactive. The algebraic, geometric, and probabilistic viewpoints of data play a key role in data mining.

Rough Sets

The theory of rough sets is motivated by practical needs to interpret, characterize, represent, and process indiscernibility of individuals. For example, if a group of patients are described by using several symptoms, many patients would share the same symptoms, and hence are indistinguishable. This forces us to think a subset of the patients as one unit, instead of many individuals.

Rough Set Theory (RST) provides a systematic method for representing and processing vague concepts caused by indiscernibility in situations with incomplete information or a lack of knowledge. At least two views can be used to interpret this theory, operator-oriented view and set-oriented view.

Rough set theory was developed by Z. Pawlak in the early 1980's. The main goal of the rough set analysis is to synthesize approximation of concept from the acquired data [1].

The philosophy of rough set is founded on the assumption that with every object of the universe of discourse we associate some information. This theory uses different approach to uncertainty. It is also used for null and missing values. The main concept of this theory is Approximation (lower and upper). The main advantage of RST is that we don't need any previous or additional information about data like probability in Statistics. This Theory overlaps with many other theories used to reasoning about data [10].

Advantage of this theory is that it allows reducing original data, to evaluate the significance of data, it is easy to understand, analyzing both quantitative and qualitative feature and also it gives straight forward interpretation of result.

In view of the available information objects characterized by the same values of the corresponding attributes are indiscernible (similar). The indiscernibility relation generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible objects is called an elementary set, and forms a basic granule of knowledge about the universe. Any union of some elementary set is referred to as crisp set otherwise the set is rough. Consequently each Rough set has boundary lines cases while crisp set have no boundary line. In the rough set approach a vague concept is replaced by a pair of well define concept called the lower and upper approximation of the vague concept. The lower approximation consists of all objects which surely belong to the concept and the upper approximation contains all objects which possible belong to the concept [17].

The difference between the lower and upper approximation constitute the boundary region consisting of the boundary line elements of the vague concept. The lower and upper approximations define the two basic operations in RST

Tele:

E-mail address: upadhyay.niharikaa@gmail.com

RST is a relatively new soft computing tool with wide range of application in many domains especially in the area of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert system, inductive reasoning and pattern recognition. The rough set approach provides efficient algorithms for finding hidden pattern in data, minimal set of data (data reduction), evaluating significance of data, generating significance of data generating sets of decision rules from data and many more [12].

Rough set based data analysis starts from a data table, called information system. The information system contains data about objects of interest, characterized in terms of some attributes. Object characterized by the same information are indiscernible (similar) in view of the available information about them.

Any set of all indiscernible objects is called an elementary set or category and forms a basic granule (atom) of knowledge about the universe [7].

Data Classifier

In this paper, five classifiers, Navies Bayes algorithm, Random Forest, k Star, Multilayer Preceptron and J48 decision tree algorithm are used for comparison. Comparison is made on accuracy.

Navies Bayesian

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems [4].

This Classification is named after Thomas Bayes, who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data [14][16].

Random Forest

Random forest is a classification and regression algorithm originally designed for the machine learning community. This algorithm is increasingly being applied to satellite and aerial image classification and the creation of continuous fields data sets, such as, percent tree cover and biomass. Random Forest has several advantages when compared with other image classification methods. It is non-parametric, capable of using continuous and categorical data sets, easy to parametrize, not sensitive to over-fitting, good at dealing with outliers in training data, and it calculates ancillary information such as classification error and variable importance [8][9].

Random forest is an ensemble model which means that it uses the results from many different models to calculate a response. In most cases the result from an ensemble model will be better than the result from any one of the individual models. In the case of random forests, several decision trees are created (grown) and the response is calculated based on the outcome of all of the decision trees [11].

K* (K Star) Algorithm

An algorithm, called K*, is used for finding the k shortest paths between a designated pair of vertices in a given directed weighted graph. It has two advantages. First, it performs on-the-fly, which means that it does not require the graph to be explicitly available and stored in main memory. Portions of the graph will be generated as needed. Secondly, it is a directed algorithm which enables the use of heuristic functions to guide the search. This leads to significant improvements in the memory and runtime demands for many practical problem instances. Its design of K is inspired by Eppstein's algorithm [3][5]. By use of K* we determine a shortest path tree T of G and use a graph structure P(G) which, as in Eppstein's algorithm, is searched using Dijkstra to determine s-t paths in the form of sidetrack edge sequences. However, as mentioned before, K* is designed to perform on-the-fly and to be guided by heuristics [6][19].

Multilayer Perceptron

The network consists of a set of sensory units (source nodes) that constitute the input one or more hidden layer of computation nodes an output layer of computation nodes. The input signal propagates through the network in a forward direction, on a layer-by-layer basis. These neural network are commonly referred to as multilayer perceptron [2][13].

The Multilayer perceptron was presented by Rumelhart and Mc Clelland in 1986. Multilayer perceptron have been applied successfully to solve some difficult and diverse problem by treating them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. This algorithm based on the error- correction Learning rules. As such, it may be viewed as a generalization of any equally popular adaptive [13][20].

J48

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tupple in the database and results in classification for that tupple [14][16].

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [18][15].

Measuring Performance

The performance of classification algorithm is usually examined by evaluating the accuracy of the classification. However since classification is often a fuzzy problem, the correct answer may depend on the user. Traditional algorithms evaluation approaches such as determining the space and time overhead can be used but these approaches are usually secondary. Determining which better best is depends on the interpretation of the problem by users. Classification accuracy is usually calculated by determining the percentage of tuples placed in a correct class. This ignores the fact that there also may be a cost associated with an incorrect assignment to the wrong class. This perhaps should also determine.

Accuracy Comparison

Accuracy Comparison for the following datasets:

- Wine dataset
- Two wheeler dataset
- Energy efficiency
- Fertility diagnosis
- Cyclic power plant
- Concrete slump test

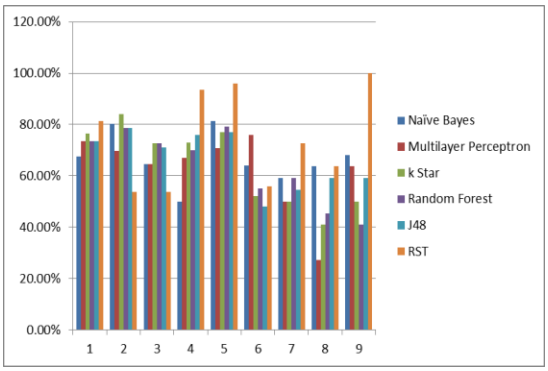
Here, firstly we find the accuracy of these dataset using all five algorithms i.e. J48, Random forest, navies Bayesian, K* and multilayer Perceptron then we find accuracy of same data set using Rough Set Theory and compare them with the accuracy.

Comparisons of Accuracy of all dataset.

Algorithm	Wine Dataset	EEHL	EECL	TWDS	FD	CPP	Slump	Flow	Strength
Naïve Bayes	67.64%	80.35%	64.51%	50%	81.25%	64%	59.09%	64%	68.18%
Multilayer Perceptron	73.52%	69.64%	64.51%	67%	70.83%	75.86%	50.00%	27.27%	63.63%
K*	76.47%	83.92%	72.58%	73%	77.08%	52%	50.00%	41%	50.00%
Random Forest	73.52%	78.57%	72.58%	70%	79.16%	55.14%	59.09%	45.45%	40.90%
J48	73.52%	78.57%	70.96%	76%	77.08%	48.00%	54.54%	59.09%	59.09%
RST	81.25%	53.85%	53.85%	93.59%	95.83%	56.00%	72.73%	63.64%	100.00%

Where,
EEHL- Energy Efficiency for Heating Load
EECL- Energy Efficiency for Cooling Load
TWDS- Two Wheeler Dataset
FD-Fertility Diagnosis
CPP-Cyclic Power Plant

Graphical Representation of Accuracy of all dataset



Conclusion

The proposed method uses various dataset. The experiments have been performed using the Weka tool and Rough Set Theory. All dataset have been taken from UCI repository. The experiments results shown in the study are about classification accuracy.

The accuracy of various datasets using different algorithms like J48, Random Forest, Navies Bayesian, Multilayer Preceptron and k* is compared by RST. The paper represents the comparison of the accuracy of all datasets using all mentioned algorithms and Rough Set Theory. Therefore RST is more efficient as compare to all algorithms, to finding accuracy.

References

1. Pawalak Z., (1982), *Rough set*, International Journal of Information & computer science, 11, 341-356.
2. Frean M., (1990), The Upstart Algorithm : A Method for Constructing and Training Feed forward Neural Networks, Neural Computation,198- 209.
3. Cleary, J. and Trigg L., (1995), *K*: An Instance-based Learner Using an Entropic Distance Measure*, 12th International Conference on Machine Learning, 108-114.
4. Kohavi R. (1996), *Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid*, Second International Conference on Knowledge Discovery and Data Mining, 202-207.
5. Chiang W.K. and Chen R.J., (1998), *Topological properties of the (n, k)-star graph*, International Journal of Foundations of Computer Science, 235–248.
6. Eppstein D., (1998), *Finding the k shortest paths*, SIAM J. Computing, 652–673.
7. Walczak B., Massart D.L., (1999), *Rough Set Theory-Tutorial*, Chemometrics and Intelligent Laboratory Systems, 1-16.
8. Breiman L., (2001), *Random forests*, Machine Learning, 5–32.
9. Liaw, A., Wiener, M., (2002). *Classification and regression by Random Forest*, R News, 18-22.
10. Zbigniew S., (2004), *An introduction to Rough set theory & its application*, ICENCO, 1-29.
11. Pal, M., (2005), *Random forest classifier for remote sensing classification*. International Journal of Remote Sensing, 217–222.
12. Rissino S. and Torres G. L., (2009), *Rough Set Theory - Fundamental Concepts, Principals, Data Extraction, and Applications* , Julio Ponce and Adem Karahoca, ISBN, 35-58.
13. Alsmadi M. K. S., Omar K, and Noah S.A., (2009), *Back Propagation Algorithm: The Best Algorithm among the Multi-layer Perceptron Algorithm*, International Journal of Computer Science and Network Security, 378-383.
14. Goyal A. and Mehta R., (2012), *Performance Comparison of Naïve Bayes and J48 Classification Algorithms*, International Journal of Applied Engineering Research, ISSN 0973-4562.
15. Robu, R., Hora, C., (2012), *Medical data mining with extended WEKA*, , 16th International Conference on *Intelligent Engineering Systems (INES)*. , 347-350.
16. Patil T.R., Sherekar S.S, (2013), *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*, International Journal Of Computer Science And Applications, 256-261.
17. Dr. Jyoti, (2013), *Rough Set Theory and its Applications*, International Journal of Innovative Research and Development, 268-271.
18. Kaur G., Chhabra A., (2014), *Improved J48 Classification Algorithm for the Prediction of Diabetes*, International Journal of Computer Applications 13-17 (WEKA).
19. Dayana C. Tejera H., (2015), *An Experimental Study of K* Algorithm*, I.J. Information Engineering and Electronic Business, 14-19.
20. Mia M.M.A., Biswas S.K., Urmi M.C., Siddique A., (2015), *An Algorithm For Training Multilayer Perceptron (MLP) For Image Reconstruction Using Neural Network Without Overfitting*, International Journal of Scientific & Technology Research, 271-275.