Florin Leon, Mihai Horia Zaharia, Dan Galea (2004)
Performance Analysis of Categorization Algorithms
Proceedings of the 8th International Symposium on Automatic Control and Computer Science, Iasi, ISBN 973-621-086-3

Performance Analysis of Categorization Algorithms

Florin Leon, Mihai Horia Zaharia, Dan Gâlea

***Abstract* — In this paper, a performance analysis for several well-known categorization algorithms is made. The main goal is to find classes of algorithms based on their error rate on the whole training set and by dividing the datasets into 2/3 for training and 1/3 for testing. Also, due to their different implementation characteristics, we find it necessary to make a comparison between their execution time when run in the same conditions. For this analysis, we use some representative benchmark problems for machine learning. Also, we introduce two new benchmark problems, starting from the observation that XOR is a typical difficult problem for most categorization algorithms.**

***Index Terms* — artificial intelligence, machine learning, categorization algorithms, comparative performance analysis and categorization benchmark problems**

I. Introduction

Living in an extremely complex environment, human beings are forced to reduce the number and diversity of stimuli in order to better control it. One of the strategies employed is *categorization*, i.e. establishing classes that include a group of objects or stimuli that have some common physical or functional traits [20]. Conversely, for most practical uses, a supervised approach is necessary, especially when using a computer-based method to assist the human user. In this case, it is important to find the rules that describe a certain category (or class). This is one of the most studied reasoning activities, as it is useful for almost every field of human knowledge. Automatic classification techniques developed by Machine Learning and Statistics researchers achieve excellent performance, given sufficient structure in the underlying relationship between characteristics and categories, and given sufficient data describing the relationship [7].

Accurate classification requires prior knowledge as to the relationship between possible categories and the patterns of feature values that will be encountered. The learning phase of classification is concerned with assembling this knowledge. Learning is complicated by the fact that data encoded from a stimulus may be insufficient to fully explain the categorization. There may be missing values or noise in the data, or the categorization may depend on features not encoded, or there may be interactions between features that have been encoded and features that have not.

Once the categorization rules have been determined, prediction is simple. Any new instance we want to classify can be placed in a uniquely established category, or more categories with different degrees of confidence or similarity.

Given the advances made in machine learning, it is essential to perform a comparative analysis on some well-known categorization algorithms with several representative benchmark problems.

II. The Algorithms

The algorithms we tested are well known in machine learning. We find it necessary to make a performance analysis comparison between them for a variety of benchmark problems, in order to provide help for a possible user who wants to select a classifier for her practical categorization problem. From the implementation point of view, we used the variants described in the data-mining book by Witten and Frank [19].

A. C4.5

C4.5 is a software extension of the basic ID3 algorithm designed by Ross Quinlan [15] to address the following issues not dealt with by ID3: avoiding overfitting the data, determining how deeply to grow a decision tree, reduced error pruning, rule post-pruning, handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, improving computational efficiency [8].

The C4.5 algorithm generates a classification-decision tree for the given dataset by recursive partitioning of data. The decision is grown using depth-first strategy [9]. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attribute.

The algorithm supports decision tree pruning at the end of the training process. In our comparison, we test both C4.5 with and without pruning. We call the C4.5 unpruned algorithm “C4.5U” (see paragraph 4).

B. Random Tree and Random Forest

Eibe Frank and Richard Kirkby proposed the random tree ("RT") classifier. It builds a tree that considers *k* random features at each node, and performs no pruning, therefore its error rate on the training set alone is rather small. Several classification trees compose a random forest ("RF") [2]. To

classify a new object from an input vector, the input vector is run down each of the trees in the forest. Each tree gives a classification, and this is considered to be a "vote" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

C. Reduced Error Pruning Tree

The Reduced Error Pruning Tree ("REPT") is a fast decision tree learner that builds a decision tree using information gain or variance reduction and prunes it using reduced-error pruning (with backfitting). It only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (as in C4.5).

D. Nearest Neighbor with Generalization

Instance-based learning reduces the learning effort by simply storing the examples presented to the learning agent, and classifying the new instances on the basis of closeness to their "neighbors", i.e. previously encountered instances with similar attribute value. Nearest Neighbor with Generalization ("NNG") is a nearest neighbor like algorithm using non-nested generalized exemplars [11]

E. PART

The PART algorithm [6] is a rather simple algorithm that does not perform global optimization to produce accurate rules; instead it adopts the separate-and-conquer strategy, i.e. it builds a rule, removes the instances it covers, and continues creating rules recursively for the remaining instances until there are no more instances left.

F. *RIPPER*

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [3] tries to increase the accuracy of rules by replacing or revising individual rules. It uses reduced error pruning, which isolates some training data in order to decide when to stop adding more conditions to a rule. It used a heuristic based on the minimum description length principle as stopping criterion. Rule induction is followed by a post-processing step that revises the rules to approximate what would have been obtained by a global pruning strategy.

G. Bayesian Network and Naïve Bayes

Bayesian rule induction is based on Bayes' theorem [1] about conditional probabilities, that determines the posterior probability distribution of X (the conditional probability distibution of X given Y), by multiplying the prior probability density function by the likelihood function, and then normalizing the result:

$$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)} \qquad (1)$$

Since there are usually more (n) attributes X_i, the formula applied is:

$$P(X_i \mid Y) = \frac{P(Y \mid X_i) \cdot P(X_i)}{\sum_{j=1}^{n} P(Y \mid X_j) \cdot P(X_j)} \qquad (2)$$

The Naïve Bayes ("NB") classifier computes the conditional probabilities of classes assuming that all attributes are independent. A Bayesian Network ("BN") is a graph associated with a set of probability tables. The nodes represent the variables, and the arcs represent causal relationships between variables

III. The Problems

In order to perform an extensive analysis of the algorithms, we considered several classic benchmark problems for machine learning and we also proposed two new ones.

A. *Iris*

This is perhaps the best-known database to be found in the pattern recognition literature. Fisher's paper [5] is a classic in the field and is referenced frequently to this day [4]. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant: Setosa, Versicolor, and Virginica.

The Setosa class is linearly separable from the other two, which in turn are not linearly separable from each other. There are 4 numeric attributes: petal length, petal width, sepal length, and sepal width, all expressed in centimeters. There are no missing attributes.

B. *Soybean*

This problem [12] tries to categorize the soybeans into 19 classes: diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury, and herbicide-injury, based on 35 categorical attributes. The attributes are discrete, some of them have Boolean values. There are attributes with missing (unknown) values.

C. *Vote*

This dataset [16] contains 435 votes from U.S. Congress in 1984 regarding 16 issues on: handicapped infants, water project cost sharing, adoption of the budget resolution, physician fee freeze, El Salvador aid, religious groups in schools, anti satellite test ban, aid to Nicaraguan contras, MX missile, immigration, synthetic fuels corporation cutback, education spending, superfund right to sue, crime, duty free exports, export to South-Africa administration act.

The problem is to identify the political orientation of a person (republican or democrat) depending on the votes (yea, nay, or unknown).

D. Mushroom

This dataset contains mushroom records drawn from "The Audubon Society Field Guide to North American Mushrooms" [10]. It includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as edible or poisonous.

There are 8124 instances, identified by 22 attributes, all with nominal values: cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root, stalk surface above ring, stalk surface below ring, stalk color below ring, veil type, veil color, ring number, ring type, spore print color, population, and habitat.

E. Breast cancer

The breast cancer ("BC") data set [13] originates from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia, provided by M. Zwitter and M. Soklić. It includes 286 instances of two classes: no recurrence events and recurrence events. There are 9 discrete attributes, either linear or nominal such as: age, menopause, tumor size, irradiated etc.

F. Labor

This dataset represents the final settlements in labor negotiations in Canadian industry. There are 57 instances defined by 16 attributes: duration of agreement, wage increase in first year of contract, wage increase in second year of contract, wage increase in third year of contract, cost of living allowance, number of working hours during week, employer contributions to pension plan, standby pay, shift differential (supplement for work on second and third shift), education allowance, number of statutory holidays, number of paid vacation days, employer's help during employee long term disability, employers contribution towards the dental plan, employer's financial contribution towards the covering the costs of bereavement, and employer's contribution towards the health plan. Some attribute are numeric, others are discrete (some of which are Boolean). There are no missing values. The classifier is supposed to predict whether the job is good or bad.

G. Arrow

The XOR function is typically difficult to learn for most algorithms. The diagonal is hard to learn for non-oblique decision tree inducers. Therefore, we proposed two benchmarks problems for categorization, based on 3 classes defined by a function of 2 normalized numerical attributes, *x* and *y*:

```
if | x² - y² | < 0.1 then type = 0
else
      if (x > 0.5) then type = 1
      else              type = 2;
      end-if
end-if
```

Instances from this distribution can be randomly generated, resulting in the image presented in figure 1. Class 0 is dark gray, class 1 is light gray, and class 2 is black.

Figure 1. The arrow problem

The second variant is based on the same distribution, but a new attribute is added: $ratio = \frac{x+1}{2 \cdot (y+1)}$. The problem tests the performance of classifiers when attributes are correlated. Also, this new attribute helps categorizing the instances from class 0, which are mainly diagonally distributed. We called this problem "arrow plus shape" ("AS").

For the analysis in paragraph 4, we randomly generated 1000 instances from both arrow problems.

H. MONK's

The three MONK's problems are another example of benchmark on which extensive studies have been performed [18]. They originated in the field of robotics, where a robot is supposed to be described by 6 attributes: head shape (round, square, octagon), body shape (round, square, octagon), whether it is smiling (yes, no), the object it is holding (sword, balloon, flag), jacket color (red, yellow, green, blue), and whether it has a tie (yes, no). The learning task is a binary classification one. Each problem is given by a logical description of a class. Robots belong either to this class or not, but instead of providing a complete class description to the learning problem, only a subset of 432 possible robots with its classification is given. The classifier must generalize over a rather small subset of these 432 examples, and predict the class membership for the instances left out.

Problem 1 is defined as: *(head shape = body shape) or (jacket color is red)*. From 432 examples, 124 were randomly selected for training, without misclassifications. The problem is in standard disjunctive normal form and should be easily learnable by symbolic learning algorithms.

Problem 2 is: *exactly two of the six attributes have their first value*. From 432 examples, 169 were randomly selected for training, without misclassifications. The problem is similar to parity problems and it combines different attributes in a way that makes it difficult to be described in DNF or CNF

using only the given attributes.

Problem 3 is defined as: *(jacket color is green and holding a sword) or (jacket color is not blue and body shape is not octagon)*. From 432 examples, 122 were randomly selected for training, with 5% misclassifications. The problem is also in DNF and serves to evaluate the algorithms under the presence of noise.

I. Drugs

This is a simple problem that tries to find the rules to prescribe one of the two classes of drugs, according to the patient's age, blood pressure and sex. The age attribute is numeric, the other two are nominal: blood pressure may be low, normal or high and sex may be female or male. A good classifier should discover that the sex attribute is irrelevant to the categorization problem. If we consider that there are two drug types: *A* and *B*, the instances are presented in table 1.

TABLE 1.
THE DRUGS PROBLEM

Sex	Age	Blood pressure	Drug
male	20	normal	A
female	73	normal	B
female	37	high	A
male	33	low	B
female	48	high	A
male	29	normal	A
female	52	normal	B
male	42	low	B
male	61	normal	B
female	30	normal	A
female	26	low	B
male	54	high	A

J. Shepard

Shepard, Hovland, and Jenkins [17] studied human performance on a category learning task involving 8 stimuli divided evenly between two categories. Varying exhaustively three binary dimensions generated the stimuli. They observed that if these dimensions are regarded as interchangeable, there are only 6 possible category structures across the stimulus set, as shown in table 2 [14]:

TABLE 2.
SHEPARD'S PROBLEMS

Stimulus	Attributes			Class					
				1	2	3	4	5	6
1	0	0	0	A	A	A	A	A	A
2	0	0	1	A	A	A	A	A	B
3	0	1	0	A	B	A	A	A	B
4	0	1	1	A	B	B	B	B	A
5	1	0	0	B	B	B	A	B	B
6	1	0	1	B	B	A	B	B	A
7	1	1	0	B	A	B	B	B	A
8	1	1	1	B	A	B	B	A	B

In the first problem only the first attribute is significant for the categorization. The second problem is a XOR-type one. Problems 3, 4, and 5 are rule-plus-exception problems (that is why we tested only problem 4, since the other two are similar). Problem 6 compels the subject to memorize all stimuli, because there is no rule to group them.

IV. Comparative Performance Analysis

A. Complex Problems

A comparative performance analysis of the algorithms on the complex problems having a large number of attributes and instances was made. The experiments were made on a computer with a Duron processor working at a frequency of 700MHz, and with 256Mb RAM.

Figures 2a and 2b graphically display the error rates of the algorithms on the benchmark problems. For each problem there are two bars: one the first shows the error rate on the training set alone, and the second shows the error rate when training with 2/3 of the dataset and testing on the 1/3 left.

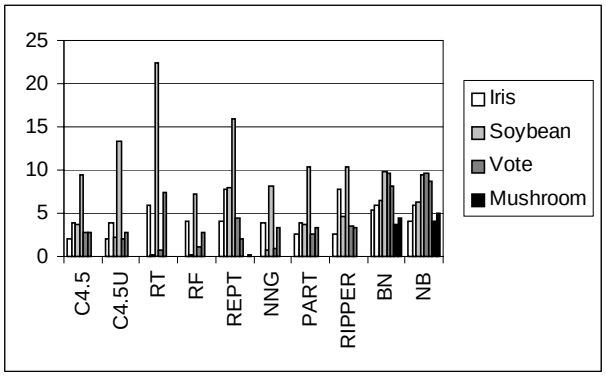

Figure 2a. The error rates on the complex problems

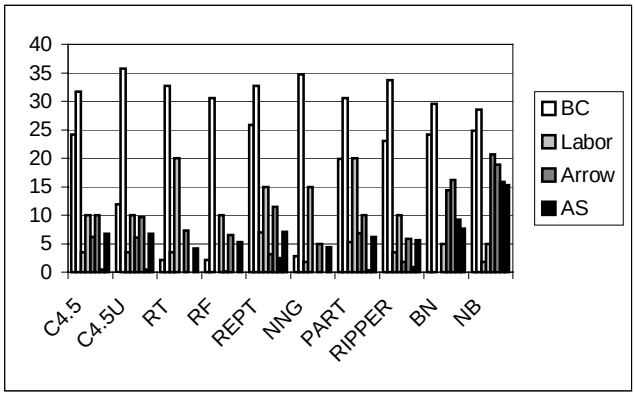

Figure 2b. The error rates on the complex problems

It seems that the Soybean, Breast-cancer, Labor and Arrow problems are the most difficult. Conversely, the Mushroom problem seems to be easily solved by most classifiers. Also Random Forest, Random Tree and Nearest Neighbor with Generalization have good performance on the training set,

they generalize rather poorly. Bayesian Network and Naïve Bayes have good generalization, but unfortunately they have a high error rate in both situations. C4.5 generally has good error rates on the training set alone (especially in the unpruned version) and acceptable generalization capabilities.

The Arrow-plus-shape problem seems to have better results than simple Arrow, as the third composed attribute helps in classifying class 0. However, one must observe the big difference in errors between the cases where testing is performed only on the training set compared to the case where 2/3 of the dataset is considered for training and 1/3 for testing. It means that most algorithms generalize very poorly on these problems. From this point of view, C4.5 performs better on Arrow than on Arrow-plus-shape. Bayesian Networks and Naïve Bayes have similar performance on the two situations, but their error rates are high nevertheless.

Figure 3 displays a comparison between the execution times of the algorithms, in seconds. In three cases, this time exceeded 3s, reaching almost 16s. However, we scaled the chart at 3s to show the performance of the faster algorithms more clearly.

Obviously, smaller problems need lesser time to be solved. However, one can notice that Random Forest and RIPPER are the slowest algorithms, while Bayesian Networks and Naïve Bayes are the fastest. PART has also good performance. C4.5 has satisfactory execution time, which is a fine indicator taking into account its good error rates.

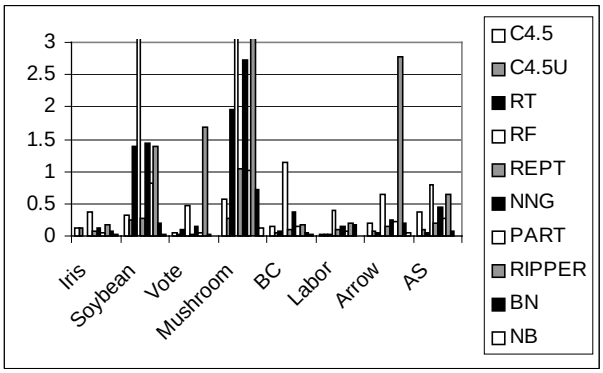

Figure 3. The execution time of the algorithms

B. The MONK's Problems

Also performance of the algorithms on the three MONK's problems was analised. The same conventions as above are used to display the data.

In this case, all the problems are similar in size, and their different difficulty levels can only explain the time difference. Figure 4 graphically shows the error rates of the algorithms we tested. On the first problem, PART, RIPPER and NNG have the best results.

On the second problem, NNG is still the best, followed by Random Forest, which however is slower. On the third problem, all the algorithms have better results, especially C4.5, Bayesian Networks and Naïve Bayes.

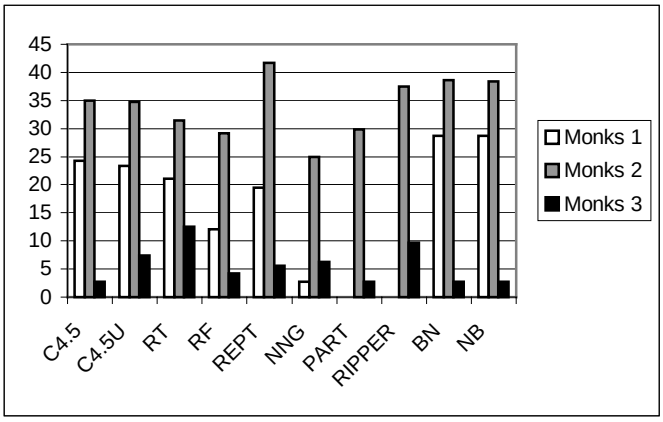

Figure 4. The error rates on the MONK's problems

C. The Logic Problems

The problems grouped in the "logic" category have an underlying set of rules that must be discovered by the classifier. Some attributes are irrelevant and should be ignored, while others can be grouped to form useful rules. Time is unsuitable for this situation, since all the problems are very small in size. Table 3 shows the performance of the algorithms for these problems.

TABLE 3.
PERFORMANCE ANALYSIS ON THE LOGIC PROBLEMS

Problem	C4.5	C4.5U	RT	RF	REPT
Drugs	0%	0%	0%	0%	0%
Shepard 1	0%	0%	0%	0%	0%
Shepard 2	50%	0%	50%	12.5%	50%
Shepard 4	25%	25%	0%	0%	25%
Shepard 6	50%	50%	50%	12.5%	50%

Problem	NNG	PART	RIPPER	BN	NB
Drugs	0%	0%	8.333%	25%	0%
Shepard 1	0%	0%	0%	0%	0%
Shepard 2	50%	50%	25%	50%	50%
Shepard 4	0%	25%	50%	0%	0%
Shepard 6	50%	50%	50%	50%	50%

The Drugs and Shepard 1 are easy, especially the latter, which poses no problem to the classifiers. Shepard 6, on the other hand, is difficult and cannot be accurately solved by any of the considered algorithms. Most of them can correctly classify only half of the instances.

V. Conclusions

Our comparison between several categorization algorithms was meant as an aid to the user who wants to choose the proper classifier for a practical purpose. From our testing, it appeared that some algorithms have a better precision but are slower, and some are faster but lack in accuracy. Also, many algorithms that have a low error rate on the training set alone, may not generalize well on previously unseen data. The results on the logical problems show that an intermediate class of categorization algorithms must be developed that can perform equally well on data drawn from natural phenomenon or chosen distributions and on formal logic problems.

REFERENCES

[1] Bayes, Rev. T., *An Essay Toward Solving a Problem in the Doctrine of Chances*, Philos. Trans. R. Soc. London 53, pp. 370-418, 1763.

[2] Breiman, L., *Random Forests*, Machine Learning 45 (1):5-32, 2001.

[3] Cohen, W. W., *Fast effective rule induction*, in Proc. Of the 12th International Conference on Machine Learning, pp. 115-123, Morgan Kaufmann, 1995.

[4] Duda, R. O., Hart, P. E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, p. 218, 1973.

[5] Fisher, R. A., *The use of multiple measurements in taxonomic problems*, Annual Eugenics, 7, Part II, 179-188, 1936.

[6] Frank, E., Witten, I. H., *Generating Accurate Rule Sets Without Global Optimization*, in Shavlik, J. (ed.), Machine Learning: Proceedings of the Fifteenth International Conference, Morgan Kaufmann Publishers, San Francisco, CA, 1998.

[7] Gâlea, D, Leon, F., et al., *Knowledge-Based Geographical Systems*, Bulletin of Technical University of Iaşi, in press.

[8] Hamilton, H., Gurak, E., Findlater, L., Olive, W., *Knowledge Discovery in Databases*, University of Regina, Canada, http://www2.cs.uregina.ca/~hamilton/courses/831/ notes/ml/dtrees/c4.5/tutorial.html, 2002.

[9] Joshi, K. P., *Analysis of Data Mining Algorithms*, http://userpages.umbc.edu/ ~kjoshi1/data-mine/proj_rpt.htm, 1997

[10] Lincoff, G. H., Knopf, A. A., *The Audubon Society Field Guide to North American Mushrooms*, pp. 500-525, 1981.

[11] Martin, B., *Instance-Based learning: Nearest Neighbor With Generalization*, Master Thesis, University of Waikato, Hamilton, New Zealand, 1995.

[12] Michalski, R. S., Chilausky, R. L., *Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis*, International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2, 1980.

[13] Michalski, R. S., Mozetic, I., Hong, J., Lavrac, N., *The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains*, in Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA, Morgan Kaufmann, 1986.

[14] Navarro, D. J., Myung, J. I., Pitt, M. A., *Comparing the Qualitative Performance of the ALCOVE and RULEX Models of Category Learning*, http://quantrm2.psy.ohio-state.edu/Navarro/alcoverulex.pdf.

[15] Quinlan, R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[16] Schlimmer, J. C., *Concept acquisition through representational adjustment*, PhD Thesis, Department of Information and Computer Science, University of California, Irvine, CA, 1987.

[17] Shepard, R. N., Hovland, C. I., Jenkins, H. M., *Learning and memorization of classiffications*, Psychological Monographs, 75, 1961.

[18] Thrun, S. B., et al., *The MONK's Problems – A Performance Comparison of Different Learning Algorithms*, Technical Report CS-CMU-91-197, Carnegie Mellon University, 1991.

[19] Witten, I. H., Frank, E., *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[20] Zaharia, M. Z., Leon, F., Gâlea, D., *On Finding an Optimized Categorization in Conceptual Spaces using Genetic Algorithms*, Bulletin of Technical University of Iaşi, in press.

With regard to performance analysis of clustering algorithms, would this be a measure of time (algorithm time complexity and the time taken to perform the clustering of the data etc) or the validity of the output of the clusters? (or both). Is there any other angle one look at to identify the performance (or lack of) for a clustering algorithm? Many thanks in advance, T. Performance Analysis of Different Clustering Algorithm. 1. Naresh MathurÂ II. Categorization Of Clustering Techniques According to Data Mining concepts and Techniques by Jiawai Han and Micheline Kamber clustering algorithm partition the dataset into optimal number of clusters. They introduce a new cluster validation criterion based on the geometric property of data partition of the dataset in order to find the proper number of clusters. The algorithm works in two stages. The first stage of the algorithm creates optimal number of clusters , where as the second stage of the algorithm detect outliers. 2.1 Cluster Algorithms: Algorithms which are being used for outlier ... K., "Performance Analysis of Various Data mining Classification Algorithms on Diabetes, Heart dataset", international journal of advanced computer technology, Vol.5 (3), 2016. Predicting Diabetes by sequencing the various Data Mining Classification Techniques. Jan 2007.Â Methods/Statistical Analysis: The impact of categorization is very important in authentic earth applications in all fields. To categorize the rudiments allowing to the applications of the elements during the predefined set of modules are used by classification methods. Very popular classification algorithms J48, Support Vector Machines (SVM), Classification and Regression Tree CART and k-Nearest Neighbor (kNN) for diabetic data are used for this research work. Classification algorithms have various practical applications in different fields such as bioinformatics, natural language processing, market segmentation and text categorization. It is used for speech recognition, facial detection, filtering spam messages, handwriting recognition, understanding spoken language, bio metric identification, document classification etc.Â Random forests or random decision forests are an ensemble learning method. It is used mainly used to solve classification, regression problems and also other problems. Random forest is one of the accurate learning algorithm. The basic concept of the algorithm is to build many small decision-tree and then merging them to form a forest.