BUILDING OPTIMAL DECISION TREES FOR CLASSIFICATION USING TAGUCHI METHOD

[1]Sridevi M., [2]ArunKumar B.R.

[1] Assistant Professor, Department of MCA, Professor & Head, Department of MCA,
BMS Institute of Technology, Bangalore, India
Email : muthyala.sridevi1@gmail.com, arunkumar.mcahod@gmail.com

Abstract – The Taguchi method is a statistical approach to improve the quality of a process by optimizing the control factors which play crucial role in determining that the output is at the target value. In this paper we combine the Orthogonal Array (OA), Signal to Noise Ratios (SNR), and the Mahalanobis Distance (MD) to build a reference point to predict the dependent variables. Knowing the highest influencing control factors, optimal decision trees are built and validated using the measures like Entropy, Gini, and Classification error.

Index Terms – Taguchi method, classification, Mahalanobis distance, Orthogonal Array (OA), SN Ratio.

I. INTRODUCTION

With the ever increasing volumes of data, it has become very essential for the organizations to collect the data, effectively deal with large amount of data and mine knowledge from the large databases. Data mining is the technique which is quite relevant in this context to extract meaningful information for the high-level management to enable them to take important decisions that affect the future of the organization. Classification is one of the striking techniques of data mining. Many techniques are being used for classification, such as Decision Trees, Linear Discriminant Analysis and Regression, Artificial Neural Networks, Nearest-Neighbour Classifier, Bayesian Classifiers, Rule-Based Classifier, etc [1]. Traditional multivariate techniques in statistics have to follow some assumptions. But it is difficult to satisfy them all at the same time. Neural network method has some drawbacks like, in providing simple clues about classifications since it gives only the birds-eye-view without having much detail. The Taguchi Method developed by Dr. Genichi Taguchi is used to improve the quality of manufactured goods by optimization which involves the best control factor levels so that the output is at the target value. Recently, it is also being applied to the fields of biotechnology, marketing and advertising, and engineering. The method is based on Orthogonal Array (OA) experiments which gives much reduced variance for the experiment with optimum settings of control parameters. "Orthogonal Arrays" (OA) provide a set of well balanced experiments and Dr. Taguchi's Signal-to-Noise ratios (S/N), which are log functions of desired output, serve as objective functions for optimization, help in data analysis and prediction of optimum results. Mahalanobis distance is used because it is a powerful method to calculate the similarity or dissimilarity of some set of conditions to an ideal set of conditions. It has a multivariate effect size [4] when compared to Euclidean distance which will take correlations of the data set into account and is also scale-variant.

The purpose of this paper is to emphasise the usage of Taguchi method to construct optimal decision trees by studying the effect of different control factors on the prediction of a dependent variable. Iris data set is taken which has a linear data structure and can be used to examine the discriminant ability of a discriminant model.

II. LITERATURE REVIEW

Classification is a task of assigning objects to one of several predefined categories. It encompasses many diverse applications in various fields. To calculate the Mahalanobis distance as response data for different combinations of the control variables, we need to define a sample “normal” observations to construct a reference value and then we identify whether the

created Mahalanobis Distance has the ability to differentiate the "normal" group from "abnormal" group. We can apply the OAs and SN ratio to evaluate the contribution of each variable and to reduce the number of variables if the number is large. In order to construct the measurement scale, we need to collect a set of ideal observations and standardize the variables of these observations to calculate the Mahalanobis distance [2].

The following is the formula for Mahalanobis distance:

$$MDj = Dj2 = (1/k)\ ZijT\ C\text{-}1\ Zij \quad (1)$$

Where Zi = standardized vector obtained by standardized values of Xi (i=1,2,...k)

$$Zij=(Xij-\bar{Xi})/Si,\ i=1,2,....k,\ j=1,2,....n$$

$\bar{Xi}$ = Mean of Xij

Where Xij = vaue of ith variable jth observation

Si = standard deviation of i-th variable

C-1 = the inverse of correlation matrix

k=number of variables

n=number of observations

T= transpose of the standard vector

OAs and SN ratios are very useful in the identification of important variables to guide in the model analysis and building decision trees in the future. Inside an OA, every run includes level combination of factors to investigate each variable's influence on the response. In the experiment design, every factor will be assigned to a column in the Orthogonal Array, and every row represents the experiment combination of a run. Each variable has two levels to represent "inclusive" and "exclusive" of the variable. We will be assigned to calculate the MD in each run, and then calculate the SN ratio from the MDs. SN ratio is defined as a tool to measure the accuracy of the measurement scale. There are 3 types of SN ratios available – larger-the-best, smaller-the-best, and nominal-the-best. We can prefer larger-the-best SN ratio because the MD in "abnormal" observations is usually larger than the MD in "normal" observations. Taguchi and Jugulum [2] also suggest using the dynamic SN. To use this, we have to know the degree of severity of each "abnormal" observation in advance. Equation (2) gives the formula for larger-the-better SN ratio-

$$SN= -10 \log [\ 1/t\ \square\ tj=1\ (1/MDj2)] \quad (2)$$

III. DEMONSTRATION

Fisher [3] used the Iris data for presenting and explaining the problems of classification. A data set with 150 samples of flowers from the IRIS species setose(species 1), versicolor (species 2), and viginica (species 3) were collected. From each species, there are 50 observations for sepal length, sepal width, petal length, petal width, and a dependent variable, i.e. the species of the flowers. Now, we need to define the "normal" observations to create a reference point. In this case, we define iris species 1 as "normal" observations, and use the reference point built from these "normal" observations to differentiate the other two different species of iris. Let us make Sepal length as factor A, Sepal width as factor B, Petal length as factor C, and Petal width as factor D as the control factors, and MD as the response variables.

Next step is to collect the "normal" observations. Table I shows three different species of iris, each with 50 samples. The data is subdivided into training samples and test samples by the ratio of 2:1. Let us randomly select 34 out of each of the 50 samples. We define species 1 as "normal" to construct the reference point. The MD of "abnormal" observation will be larger than the MD of "normal" observation. We use a test sample to validate and calculate the MD for each observation. The MDs range of species 1 is 0.087-4.713; for species 2 it is 60.371-153.615 and for species 3 it is 145.879-329.744. The result states that the reference point constructed by all 4 variables is accurate because it clearly differentiates these three different kinds of flowers. We can easily identify species 1 from the other two species of iris with 100% accuracy but there is a slight overlap between species 2 and species 3, which may require the establishment of a threshold to differentiate them.

Table I. Iris Samples

Species of iris	**Species 1**	**Species 2**	**Species 3**
Training sample	34	34	34
Testing sample	16	16	16
Total	50	50	50

Using $L_8 (2^7)$ OA and SN ratio, important variables are selected [5]. For this we randomly select 5 samples each from species 2 and species 3 to perform the analysis. Table II shows the OA allocation, the MDs and the SN ratio for each

Table II. $L_8(2^7)$ OA and SN ratios of the iris samples

Run	1(A)	2(B)	3(C)	4(D)	5	6	7	Mahalanobis Distance (MD)										SN
1	1	1	1	1	1	1	1	205.68	248.62	288.13	180.26	145.88	121.38	115.34	146.47	153.62	78.653	42.8124
2	1	1	1	2	2	2	2	180.38	298.7	354	229.94	147.66	147.83	130.7	146.36	177.63	86.926	43.6682
3	1	2	2	1	1	2	2	258.4	184.14	183.58	81.686	149.5	80.266	94.885	152.98	114.26	67.894	40.5983
4	1	2	2	2	2	1	1	24.728	71.45	61.981	10.45	12.41	27.697	19.296	7.037	8.66	10.45	21.91
5	2	1	2	1	2	1	2	276.21	192.01	206.34	95.516	163.45	92.745	108.54	157.51	127.09	78.209	41.6902
6	2	1	2	2	1	2	1	1.057	0.997	2.355	4.167	2.355	2.355	2.355	0.272	3.197	2.355	-2.1681
7	2	2	1	1	2	2	1	349.94	450.53	496.37	297.83	240.9	204.75	193.06	240.9	250.12	129.52	47.2316
8	2	2	1	2	1	1	2	477.8	831.89	935.65	584.99	381.45	381.45	337.34	381.45	452.7	221.26	51.8975

sample. Here, "1" represents including the variable and "2" represents excluding the variable. The larger-the-better ratio is applied to conduct the analysis.

Table III shows the response table for Signal to Noise Ratios for larger-the-better SN ratio and the influence that each control factor has on the prediction. Depending on the rank, C that is the petal length has the highest influence in predicting the species followed by D, petal width and B, the sepal width. So these are the important variables for predicting the species.

Table III. Response table for SN Ratios

Response Table for Signal to Noise Ratios				
Larger is better				
Level	**A**	**B**	**C**	**D**
1	37.25	31.50	46.40	43.08
2	34.66	40.41	25.51	28.83
Delta	2.58	8.91	20.89	14.26
Rank	4	3	1	2

Fig 1 represents the main effects of control variables on the prediction process. From the figure the explanation given above is obvious.

From the above result, we can attempt to draw a decision tree with the highest ranked control factor as the root node or the best split point to start the decision tree as it has the capacity of splitting the test records unambiguously and with pure partitions. The next split point can be the next ranked control factor. By this time, if all the records have identical attribute values or all the records belong to the same class the tree expansion can be stopped. Otherwise, the next best split point can be the next ranked control factor and so on.

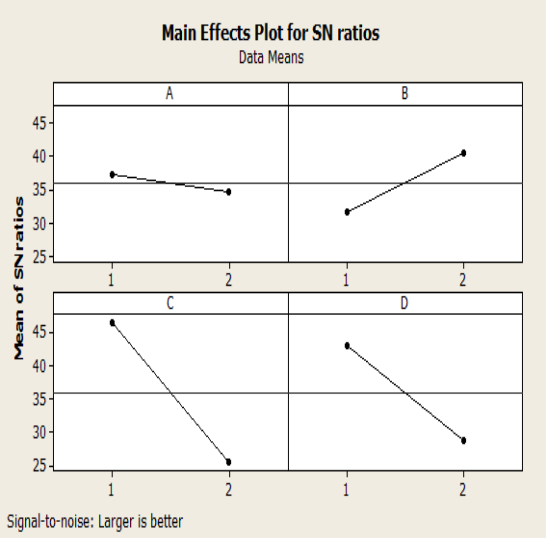

Figure 1. Main effects plot for SN ratios

The best split can be validated by using the following impurity measures [1]–

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i/t)\log_2 p(i/t) \quad (3)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i/t)]^2 \quad (4)$$

$$\text{Classification Error}(t) = 1 - \max_i [p(i/t)] \quad (5)$$

Hence, by using the Taguchi method, finding the optimal decision tree has become computationally feasible irrespective of the exponential size of the search space.

IV CONCLUSION

Optimization of decision trees in the area of data mining is a key research issue as large volumes of heterogeneous data has to be mined as per data requirement. Taguchi method is a very efficient and successful statistical analysis approach in the area of material science research. This research work introduces Taguchi method as a novel approach for building optimized decision trees. The decision trees are computed based on rank of the control factors generated by applying the Signal to Noise ratio by following Taguchi methodology.

REFERENCES

[1] Pang-Ning Tan, Micheal Steinbach, and Vipin Kumar, "Introduction to Data Mining", Pearson Education, 2006.

[2] Taguchi, G. and R. Jugulum, "New Trends in Mutivariate Diagnosis," Sankhyaa: The Indian Journal of Statistics, 62, Series B, 2, 233-248(2000).

[3] Fisher R. A., "The use of multiple measurements in taxonomic problems," Annals of Eugenics, 7, 179-188(1936).

[4] Johnson R. A. And Wichern D. W., Applied Multivariate Statistical Analysis, McGraw-Hill Press, New York (1998).

[5] Ranjit K. Roy, "Design of Experiments using the Taguchi Approach", Wiley Publications, 2001.