"BOF" Trees Diagram as a Visual Way to Improve Interpretability of Tree Ensembles

Vesna Lužar-Stiffler[1,2]
Charles Stiffler[2]
[1]*University of Zagreb, University Computing Centre*
[2]*CAIR Research Centre, Zagreb, Croatia*
vluzar@srce.hr, charles.stiffler@cair-center.hr

Abstract. *The motivation for this research stemmed from a desire to create visual aids to help researchers/managers interpret ensembles of decision tree outputs generated by various algorithms. The method employed a simulation experiment (using only bagging) followed by application of the new visualization tools on actual survey data. Simulated data, with a pre-specified structure, were "bagged" with the results presented using five graphical tools that recreated (and/or portrayed) the known data structures captured by the bagging algorithm. Then the same methodology was generalized to a structurally unknown, virgin (survey) data set. Results of the research are that five visual aids tools were examined (two of which are new approaches) and found to be useful for making action oriented interpretations from e.g., web-survey data.*

Keywords. Classification trees, BOF, bagging, visualization tools, web-survey, tree ensembles, data mining.

1. Introduction

Great deal of current classification (decision) tree research focuses on improving predictability by using various aggregation-based approaches ([1],[2],[3],[8],[14], etc). However, there is much less information available that addresses the issue of "ensemble" interpretability. The decision to undertake this research was prompted by a desire to create visual aids for interpreting ensembles of decision tree outputs.

Classification trees were first introduced by the statistical community [12], [5] and subsequently developed and popularized by researchers in the areas of machine learning / computer science ([13], etc.). Among the most well-known tree algorithms are CART [5], C4.5 [13] and CHAID [10].

Despite their many positive features (ability to handle data of mixed type and missing values, robustness to outliers, etc.), classification trees have the one major disadvantage of being unstable vis-à-vis seemingly minor data perturbations (e.g., sampling), thus lowering their predictive power. One area of improvement has been proposed in the way of ensembles of trees obtained from bootstrap [7] re-samples. Among these are Breiman's "bagging" trees [2], and the more recently introduced "Random Forests" [3]. Other ensemble algorithms, in which a large number of tree classifiers are "trained" on the training set and then combined to provide an improved aggregate / ensemble classifier, include AdaBoost [8], stacking [14] randomized trees [1], etc. Some of these combinations use equal weighting (e.g., bagging), whereas in other tree predictor scenarios (e.g., boosting) weights are subsequently adjusted.

(Note: Here we demonstrate the visualization method using bagging outputs; however it applies as well to the many other ensemble techniques.)

As mentioned, prediction error has been reduced by various methods; however, in most cases the improvement comes at the expense of interpretability, i.e., user/manager confidence in explaining, planning, and making specific action oriented decisions based on the ensemble tree outputs is severely hindered. For example, when a decision tree algorithm is applied to, say… three successive random samples of (equal) size n from a large data set produce different subsets of variables purporting to interpret "churn" (or retention, up-sell/cross-selling, or fraud, or loan default) managers are rightfully confused.

The following research offers a procedure that retains the benefit of improved predictability provided by bagging, etc., while returning the benefit of researcher/operational interpretability.

The basic research methodology was to use simulated data with a predetermined (i.e., known) structure, apply bagged trees and then present the results using 5 different types of display tools designed/selected to "recreate" (or rediscover) the structure inherent in the data and captured by the bagging algorithm. Each display

is intended to identify additional patterns in the data/algorithm and thus improve interpretability. We then apply the same approach to a "real" data set (web survey) and demonstrate the interpretational benefits of the various proposed visualization tools.

The paper is organized along the following lines: In Section 2 we briefly explain the basic idea behind the bagging ensemble technique and the proposed set of graphical displays. The simulation experiment and resulting tree outputs are displayed and explained in Section 3. In Section 4, we apply the proposed methodology to data obtained from a web survey concerning ICT usage in the Croatian primary and secondary school system. Conclusions are offered in Section 5.

2. Ensemble of trees and visualization

As mentioned previously, for simplicity and without loss of generalization, we'll limit our current experiment to "bagging classification trees" [2], which was one of the first in a series of aggregation-based tree models introduced over the period of the last ten years.

Suppose that our data arose from a (general) statistical (learning) model (i.e., data mining model)

$$Y = f(X) + \varepsilon,$$

where the random error ε has E(ε)=0, and is independent of X, where X are predictors, and Y is a response variable.

For purposes of this research, we assume that Y is restricted to 0/1 values (i.e., the Y variable is the result of some Bernoulli process).

The goal then of statistical learning would be to find a useful approximation

$$\hat{f}(x)$$

to the function $f(x)$.

2.1 Ensemble trees: bagging

Let

$$\hat{f}(x)$$

be the classification tree prediction at input x obtained from the full "training" data $Z=\{(x_1,y_1),(x_2,y_2)\ldots(x_N,y_N)\}$

Let

$$\hat{f}^{*b}(x)$$

be the classification tree prediction at input x obtained from the bootstrap sample Z*b, b=1,2,…B.

The bagging estimate is defined by

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

The above aggregation can be implemented either as "majority-rule voting" (i.e., the predicted class is the one with the most "votes" from the B trees), or as averaging class probabilities over the B trees.

It has been shown [2],[9] that bagging can lead to improved prediction by reducing variance. Other tree ensembles (e.g., Random Forests) can reduce both variance and bias. However, here we are concerned only with the issue of providing tools for qualitative understanding of the relationship between the input (predictor) variables and the resulting responses (i.e., primarily with interpretability).

2.2 Visualization

In most tree applications, as in other data mining applications, predictor variables are not equally relevant. Often, especially given a large number of predictors, only a few variables have substantial influence on the response. The relative importance of a predictor variable X_k for a single decision tree was introduced by Breiman et.al. [4] in 1984, and described in [4] as:

$$\hat{I}_k^2(T) = \sum_{t=1}^{J-1}\hat{i}_t^2 I(v(t)=k),$$

where the sum is over the J-1 internal nodes for which X_k was chosen as the splitting variable. At each node t, one of the input variables $X_{v(t)}$ is used to partition the region associated with that node into two sub-regions; within each a separate constant is fit to the response values. The particular variable chosen is the one that gives maximal estimated improvement $\hat{i}_t^2$ in squared error risk over that for a constant fit over the entire region.

In the case of bagged trees, the importance is just the average over B trees:

$$\hat{I}_k^2 = \frac{1}{B}\sum_{b=1}^{B}\hat{I}_k^2(T_b)$$

In conjunction with Random Forests, Breiman [3] introduced another measure of relative importance, based on his "out of the bag" concept, which shows promising results. Here we show only the results based on the "standard" measure of importance, as is currently estimated by SAS Enterprise Miner[1] software.

[1] SAS is a registered trademark of SAS Institute Inc. in the USA and other countries.

While the first three (of five) graphical displays (discussed below) are based on the predictor's importance measure, the last two are based on a measure of the proximity of cases (i.e., observations), and were introduced by Breiman for the visualization of Random Forests outputs.

The proximities are obtained using the following algorithm:

1. Repeat for b =1 to B: Apply the tree T_b to the training set. If case i and case j both "land" in the same terminal node, increase the proximity between i and j by 1.
2. Divide proximities by B, and set the proximity between case and itself to 1. Form an n x n (n = number of cases in the training set) proximity matrix P.

In the rest of this Section we briefly explain each of the proposed diagrams.

Diagram 1 ("Mean Importance") is a simple bar chart of averaged importance measures for all predictor variables.

Diagram 2 ("BOF Clusters") is the cluster means chart showing clusters of "similar" trees formed (or that visually flocked together, like Birds of a Feather – BOF) from the B x p matrix F of individual importance measures of p predictor variables "rated" by B trees T_b (b=1 to B).

Diagram 3 ("BOF MDPREF") is the multidimensional preference bi-plot [6] based on singular value decomposition of the F matrix. The tree vector points in (approximately) the direction of the tree's most preferred (important) variables (points), with preference increasing as the vector moves away from the origin.

Diagram 4 ("Proximity Clusters") is the cluster means chart showing clusters of "similar" cases formed from the matrix of proximities P, as "rated" by B trees.

Diagram 5 ("Proximity MDS") is the multidimensional scaling plot of "similar" cases formed from the matrix of squared distances D (D=I-P) between the cases, as "rated" by B trees.

Partial dependence plots, discussed by Hasti, et. al. [9] are an alternative, potentially useful visualization tool for ensembles, but because they require data sets with a larger number of cases, we did not apply them in our simulation experiments.

In the next two Sections we demonstrate the use of Diagrams 1-3, and either 4 or 5 for both simulated and real data.

3. Simulation Experiments

The first simulated data set (S1) uses an example introduced by Hasti et. al. [9] for the purpose of demonstrating the test error rate reductions made possible by using the bagged trees technique. It can be defined as follows:

Generate a sample of size n=30, with two classes and p=5 variables (x_1-x_5), each having a standard normal distribution with pair-wise correlation 0.95.

The responses are generated according to $Pr(Y=1|x_1 \leq 0.5) = 0.2$, $Pr(Y=1|x_1 > 0.5) = 0.8$.

(The Bayes error is 0.2.)

A test sample of size 2000 was also generated.

Classification trees (CART algorithm) were fit to the training sample and to each of B=100 bootstrap samples. (Pruning was not used.)

The second simulated data set (S2) differs from the first: the pair-wise correlation remains at 0.95 between x_1 and x_2; however, the other pair-wise correlations are set to 0.

At first glance, the diagrams in Figures 1 and 2 "Mean Importance" for data sets S1 and S2, presented in Figures 1 and 2 respectively, do not reveal much differentiation between the two datasets. Closer examination indicates that the decrease in average importance measures for the S2 data set is more nonlinear than it is for S1 (as is expected).

Still, the value of this simple chart is questionable given the small number of cases and variables in the case of S1 and S2. It's value will be demonstrated more convincingly on the real data using a larger number of cases and many more variables.

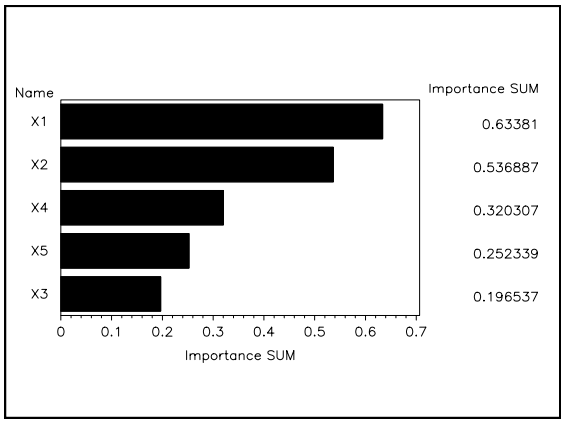

Figure 1. Diagram 1, Mean importance for the simulated data set S1, n= 30

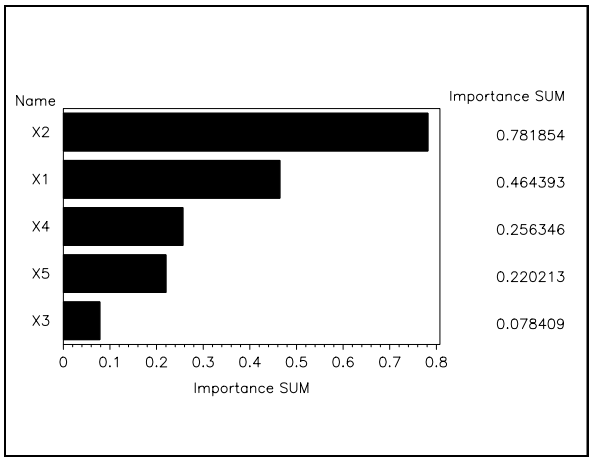

Figure 2. Diagram 1, Mean importance for the simulated data set S2, n= 30

Figures 3 and 4 are show much more differentiation among the structures "rediscovered" from the two data sets. From data set S1 BOF identified three clusters of "similar trees": Cluster 1 encompasses trees in which (on average) only predictor x_1 was "rated" as being "important" (for building the trees). In cluster 2 are trees that split mostly on x_1 and x_3; in cluster 3 on x_1 and x_4. This unstable selection of predictors seems to reflect the multicollinearity introduced for the simulated data experiment.

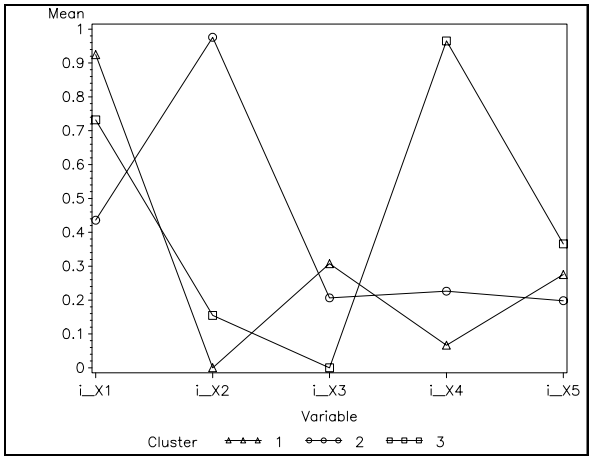

Figure 3. Diagram 2, "BOF Clusters" for the simulated data set S1, n= 30

On the other hand, Diagram 2 ("BOF Clusters") for data set S2 (Figure 4) shows almost perfect "recreation" of the induced pattern of "surrogate" variables…the first cluster contain trees that split almost exclusively on x_1, the second on x_2, with cluster 3 splitting on both x_1 and x_2.

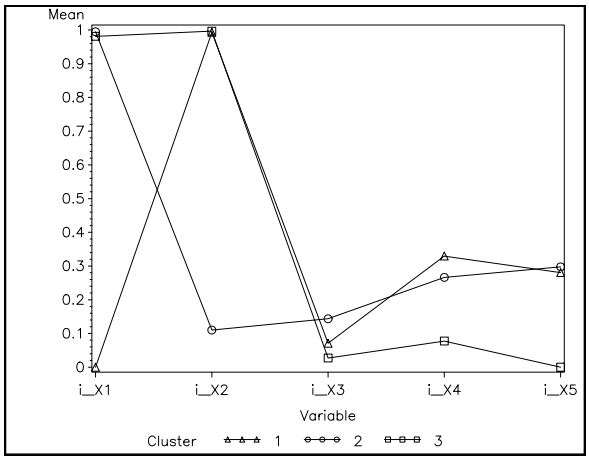

Figure 4. Diagram 2, "BOF Clusters" for simulated data set S2

These same findings can be confirmed using Diagram 3 ("BOF MDPREF"), shown in Figures 5 and 6. Clearly, there is not much "preference" shown for any particular predictor variable in S1, while there are obvious preferences for x_1, x_2, and both x_1 and x_2 in S2.

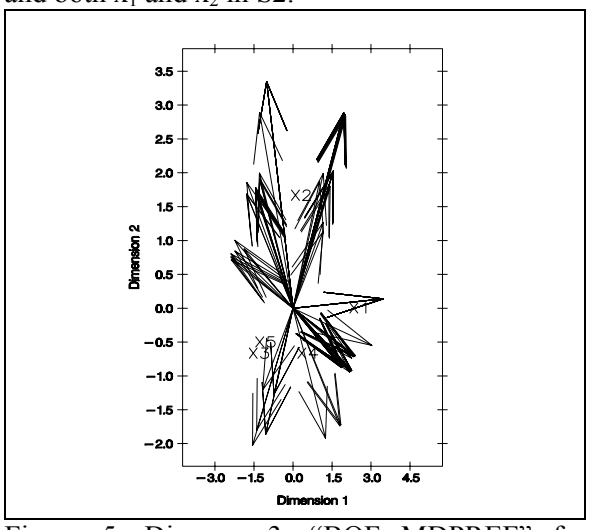

Figure 5. Diagram 3, "BOF MDPREF" for simulated data set S1

Figure 7 portrays clustering of similar cases as captured by the bagged tree algorithm applied to the S1 data set. Cluster one includes cases with the centroid located on a negative pole (for all predictor variables) and includes only the cases that were classified as 0. Cluster two is at the other extreme (i.e., all predictor variables means being positive and close to 1) and includes approximately 80% of the cases classified as 1 in this cluster. Cluster three's variable means are slightly above zero, with the overall average close to 0.5, containing approximately 20% of the cases classified as 1 in this cluster.

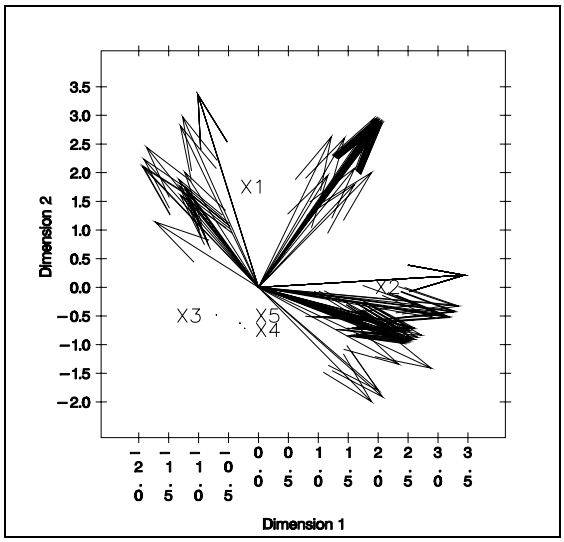

Figure 6. Diagram 3, "BOF MDPREF" for simulated data set S2

By this (BOF) method, the known data structure can be inferred (or recreated in its essence) almost perfectly. BOF gives us a panoramic view of the forest of trees being generated by the bootstrap.

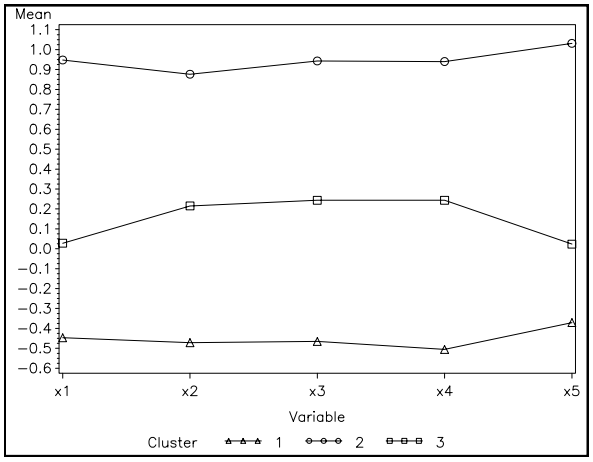

Figure 7. Diagram 4, "Proximity Clusters" for the simulated data set S1, n= 30

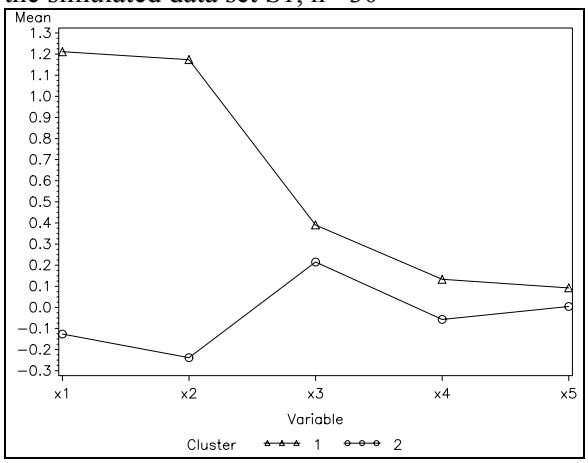

Figure 8. Diagram 4, "Proximity Clusters" for the simulated data set S1, n= 30

Finally, the same approach ("Proximity Clusters" diagram), when applied in the case of bicollinearity (S2) produces 2 clusters: one with high positive x_1 and x_2 means and almost all cases classified as 1, and the other with all cluster means relatively low, and with approximately 20% of the cases classified as 1.

4. Survey Data

The proposed visualization tools were applied to data from a recent survey of ICT usage in Croatian primary and secondary schools [11]. Here we present the results of the bagged tree algorithm applied to 200+ variables and 25,000+ cases (partitioned into training, validation and test sets in a 50%: 25%: 25% ratio). The response variable that was of major interest was a 0/1 variable "classroom use of a computer by educators".

The value of diagram 1 is obvious in this case: only the first 4 or 5 variables are relevant to explaining the targeted behavior of the educators.

A more insightful view is provided with diagrams 2 and 3: (BOF Trees): In the first group (cluster) are trees that start by splitting on variable q038, in the second group, splitting starts with Q039, and in the third cluster are trees that start with either Q024, Q038, Q046 or Z8. (Q038 and Q039 seem to be surrogate variables.)

Diagram 5 ("Proximity MDS") shows 3 groups of observations: A group of educators who never use a computer in the classroom, and the other two clusters, differentiated along the second dimension. Additional interpretably useful information can be provided by examining the correlations among MDS dimensions 1 and 2 and the predictor variables.

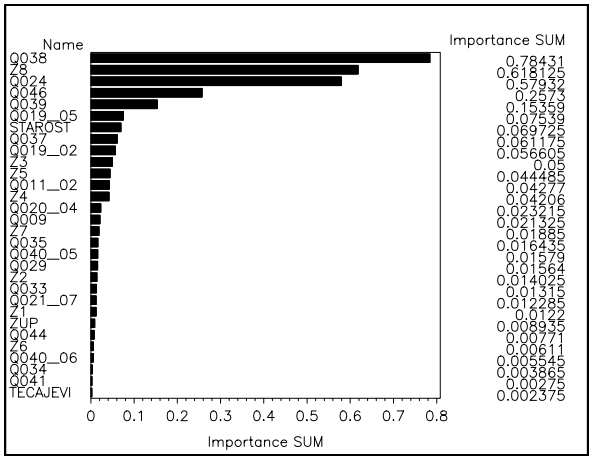

Figure 9. Diagram 1, "Mean importance" for the survey data

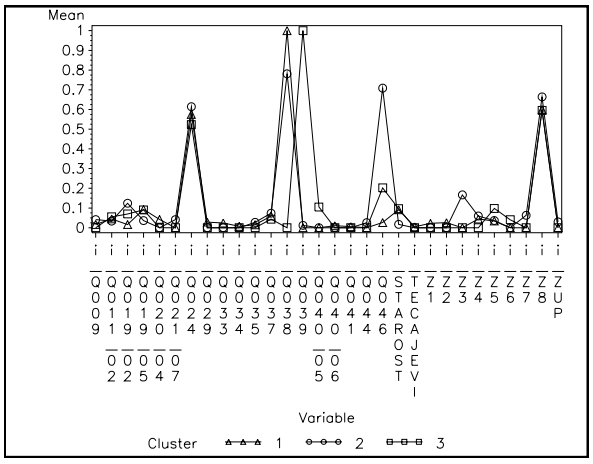

Figure 10. Diagram 2, "BOF Clusters" for the survey data

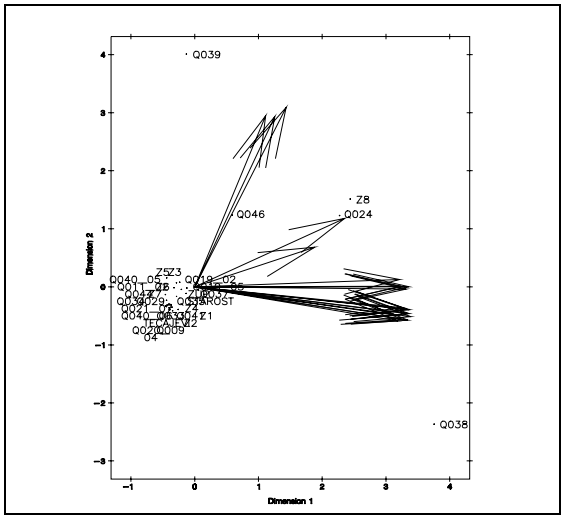

Figure 11. Diagram 3, "BOF MDPREF" for survey data

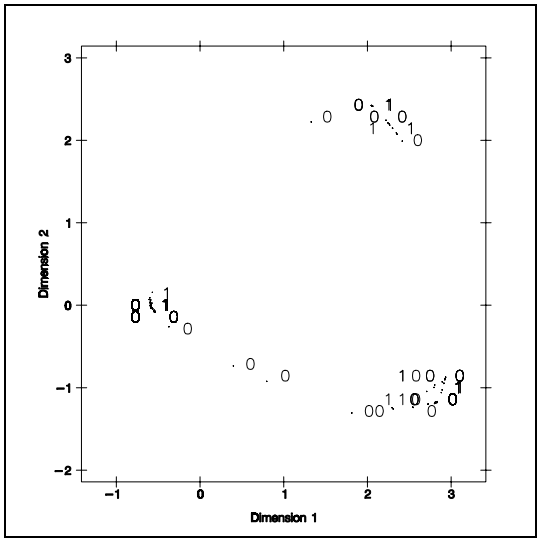

Figure 12. Diagram 5, "Proximity MDS" for the survey data

5. Conclusion

A useful "suite" of visualization tools (five altogether, with two being original inventions / applications) has been examined and applied to both simulated and actual date structure problems/ predictions.

Results indicate that since BOF type graphics more easily enable action oriented interpretation, software development effort could profitably be applied to "seamlessly" linking outputs from ensemble algorithms directly to the types of visual "information technology interface" tools presented above.

Future research is envisioned for expansion into multi-class problems (versus the binary responses used above) and into wider Monte Carlo experimentations (different pre-defined structures, different variable / sample sizes, etc.).

6. References

[1] Amit Y. and Geman D. Shape Quantization and recognition with randomized trees. *Neural Computation*, 9, 1545-1588, 1997.

[2] Breiman L. Bagging Predictors. *Machine Learning*, Vol 26, 123-140, 1996.

[3] Breiman L. Random Forests. *Machine Learning*, Vol 45, xx - xx, 2001.

[4] Breiman L. Wald Lecture II: Looking inside the black box

[5] Breiman, L., Friedman, J., Olshen, R. and Stone C. *Classification and Regression Trees*. Wadsworth, 1984.

[6] Carroll, J.D. Individual differences and multidimensional scaling, Multidimesional Scaling: Theory and Applications in the *Behavioral Sciences*, Vol.1, Ed. Shepart, R.N., Romney, A.K., and Nerlove, S.B., Seminar Press, 1972.

[7] Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26, 1979.

[8] Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55(1), 119-139, 1997.

[9] Hasti T., Tibshirani R., Friedman J. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction.* Springer Verlag, 2001.

[10] Kass G.V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29, 119-127, 1980.
[11] Luzar-Stiffler V., Stiffler C. A survey of primary and secondary school ICT infrastructure status/utilization, educator skills/knowledge, and future training needs, CARNet / Srce Project Report, 2004.
[12] Morgan J., Sonquist J.A. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 415-435, 1963.
[13] Quinlan R. C4.5: *Programs for Machine Learning*, Morgan Kaufman, San Mateo, 1993.
[14] Wolpert D. Stacked Generalization, *Neural Networks*, 5, 241-259, 1992.