# Attribute Space Extension for Accuracy Improvement

**Authors: Md Nasim Adnan and Md Zahidul Islam**
**Affiliation: School of Computing and Mathematics**

## Abstract

Typically the prediction accuracy of a decision tree/forest increases when they are built from the extended space dataset. Often attribute space is extended by randomly combining two or more attributes. In this poster we propose some novel approaches for the space extension where we only choose those attributes that have high classification capacity from the set of all attributes generated from combination.

## Motivation and Problem Statement

M. F. Amasyali and O. K. Ersoy [1] presented detailed experimental results on the effect of attribute space extension for some forest algorithms. It appeared that all the forests delivered better ensemble accuracy when applied on extended attribute space. Interestingly, Bagging, Random Subspace and Random Forest achieved higher individual accuracy of trees when they were applied on extended space. However, the study considered only one method of attribute space extension (by combining two randomly selected attributes). Thus the possible method of attribute space extension with the best outcome remains unclear.

## Proposed Attribute Space Extension Technique

**Step 1: Combine existing attributes to produce a set of candidate attributes.**

The new attributes are generated by combining $k$ number of exiting attributes. Therefore, we get a set of candidate attributes $\boldsymbol{A}_C$, where the size of the set is $|\boldsymbol{A}_C| = {}^dC_k$. Note that our technique is not restricted to any specific $k$ value. In literature, $k$ is selected as 2 [1].

**Step 2: Select new attributes from the set of candidate attributes.**

We next compute the Gain Ratio [2] of each of the ${}^2C_k$ number of candidate attributes. Based on the Gain Ratio, we select the best $d'$ attributes ($\boldsymbol{A}_{d'}$) from ${}^2C_k$. The selected $d'$ attributes are then added in the original data set. Therefore, we get the extended attribute space $\boldsymbol{A}_E = \boldsymbol{A}_O \cup \boldsymbol{A}_{d'}$. Our proposed technique is not restricted to any specific $d'$ value. In literature different $d'$ values (such as $d$, $2d$, and $3d$) have been tested [1]. However, we experiment on two different $d'$ values: $d/2$ and $d$.

## Applications of the Proposed Technique

1. Increasing Accuracy for C4.5 Trees (see Table 2 and Figure 1).
2. Increasing Accuracy for Random Forest (see Table 3).

In this experimentation, we select five (5) different data sets (see Table 1) that are publicly available from the UCI Machine Learning Repository [3].

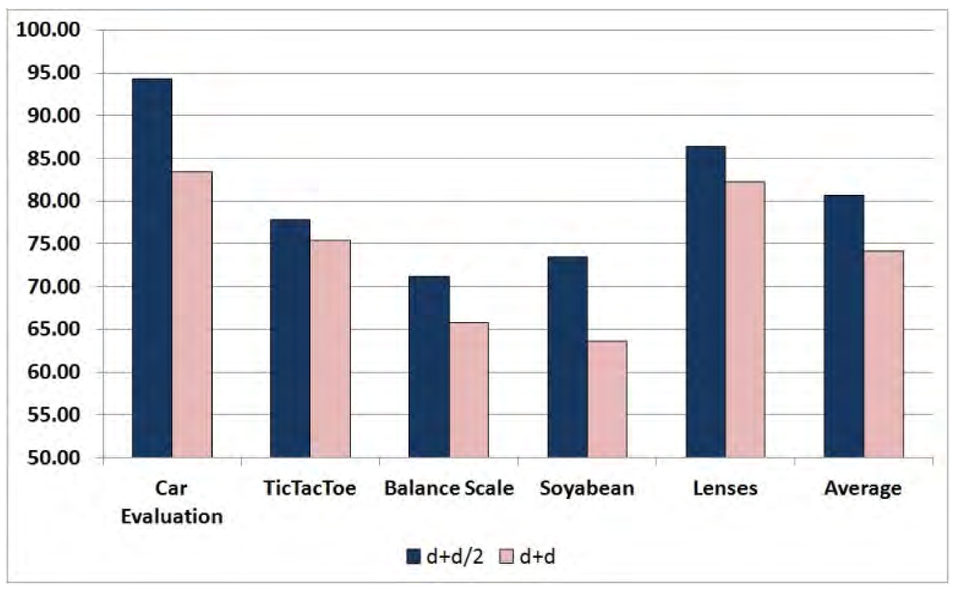
### Table 1

| Data Sets | Attributes | Number of Records | Number of the Class Attribute |
|---|---|---|---|
| Car Evaluation | 6 | 1728 | 4 |
| Tic-Tac-Toe | 9 | 958 | 2 |
| Balance Scale | 4 | 625 | 3 |
| Soybean | 35 | 47 | 4 |
| Lenses | 4 | 24 | 3 |

### Table 2

| Dataset Name | C4.5 on $A_O$ | C4.5 on $A_E$ | C4.5 on $A_{rnd}$ |
|---|---|---|---|
| Car Evaluation | **94.0950** | 93.9250 | 93.3020 |
| Tic-Tac-Toe | 82.5850 | **83.5170** | 78.4880 |
| Balance Scale | 65.6000 | **70.4490** | 67.6350 |
| Soybean | 72.2730 | **72.9550** | 62.9550 |
| Lenses | 83.3330 | **86.6670** | 80.0000 |
| **Average** | 79.5772 | **81.5026** | 76.4760 |

### Figure 1




### Table 3

| Dataset Name | RF on $A_{O,d}$ | RF on $A_{E,d/2}$ | RF on $_{E,d}$ | RF on $A_{rnd,d}$ |
|---|---|---|---|---|
| Car Evaluation | **92.9340** | 92.3030 | 92.0680 | 88.5520 |
| Tic-Tac-Toe | **92.8100** | 90.2840 | 86.1480 | 84.6500 |
| Balance Scale | 78.7170 | **88.8420** | 80.1800 | 88.1730 |
| Soybean | 97.5000 | **100.0000** | 97.5000 | 95.0000 |
| Lenses | 78.3330 | **86.6670** | **86.6670** | 50.0000 |
| **Average** | 88.0588 | **91.6192** | 88.5126 | 81.2750 |

## References

[1] M. F. Amasyali, and O. K. Ersoy, “Classifier Ensembles with the Extended Space Forest”, IEEE Transaction on Knowledge and Data Engineering, vol. 26, pp. 549-562, March 2014.

[2] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers Inc., San Francisco, U.S.A., 1993.

[3] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html (Last Accessed: 15 Oct 2015)

**Contact details:**

**Md Nasim Adnan**

Phone: +61 431920651

Email: madnan@csu.edu.au