Fault Diagnosis in Benchmark Process Control System Using Stochastic Gradient Boosted Decision Trees

Tarun Chopra, Jayashri Vajpai

Abstract—**Decision trees create an easily understood structure for evaluating complex decisions. Tree Boost models often have a degree of accuracy that cannot be obtained using a large, single-tree model. Tree Boost models are adaptable, easy to interpret and often equal to or superior to any other predictive functions including neural networks. In this paper, the performance of the proposed approach based on Stochastic Gradient Boosted Decision Trees based method is demonstrated on the DAMADICS benchmark problem. An attempt has been made to improve the performance of fault diagnosis task on DAMADICS benchmark.**

Index Terms— **Fault Diagnosis, Stochastic Gradient Boosted Decision Trees, DAMADICS**

I. Introduction

Decision tree is a hierarchical tree structure which is used to classify data on the basis of a series of rules about the attributes of the underlying classes. Recent advancements in decision tree analyses include the Tree Boost method developed by Jerome Friedman [1].Tree Boost algorithm is optimized for improving the accuracy of models built on decision trees. This method use ensembles of trees to increase the predictive accuracy over a single-tree model. Tree Boost is also known as “Stochastic Gradient Boosting” and “Multiple Additive Regression Trees”. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized. In many cases, the predictive accuracy of such a series greatly exceeds the accuracy of the base function used alone.

The aforementioned merits of tree boost technique motivated the authors to adopt this technique for decision making in relation to Fault diagnosis in a Complex Benchmark Process Control System, with multiple measured variables and overlapping fault classes.

II. STATE OF ART

Many classification models have been proposed in the literature [2]. Decision trees are especially attractive for a data mining environment for three reasons. First, due to their intuitive representation, they are easy to assimilate by humans [3]. Second, they can be constructed relatively fast compared to other methods [4]. Last, the accuracy of decision tree classifiers is comparable or superior to other models [5].

Since decision trees were introduced by Qinlan [6], they have become a highly successful learning model and are used for both classification and regression. Friedman furthered the usage of decision trees in machine learning with the introduction of stochastic gradient boosted decision trees [GBDT], using regression trees as weak learners.GBDT is also highly adaptable and many different loss functions can be used during boosting. More recently, adaptations of GBDT utilizing pair wise and ranking specific loss functions have performed well at improving search relevance [7-8]. In addition to its advantages in interpretability, GBDT is able to model feature interactions and inherently perform feature selection. Besides utilizing shallow decision trees, trees in stochastic GBDT are trained on a randomly selected subset of the training data and are less prone to over-fitting [9].

These features have made Stochastic GBDT, one of the most widely used learning algorithms in machine learning today.

III. TREE BOOST APPROACH

Stochastic GBDT is an additive regression model consisting of an ensemble of regression trees.

Mathematically, a Tree Boost model can be described as:

Predicted Target = F0 + B1*T1(X) + B2*T2(X) + … + BM*TM(X)

Where:- F0 is the starting value for the series (the median target value for a regression model), X is a vector of pseudo-residual values remaining at this point in the series, T1(X), T2(X) are trees fitted to the pseudo-residuals, &B1, B2, etc. are coefficients of the tree node predicted values that are computed by the Tree Boost algorithm.

Graphically, a Tree Boost model can be represented as shown in fig1:

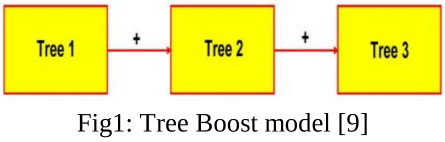

Fig1: Tree Boost model [9]

The first tree is fitted to the data. The residuals (error values) from the first tree are then fed into the second tree

Manuscript received June 26, 2011.

Tarun Chopra, Ph.D. Scholar, Department of Electrical Engineering, Faculty of Engineering, J.N.V. University, Jodhpur (India)-342 011 (email: tarun_ecb@rediffmail.com)

Dr. Jayashri Vajpai, Associate Professor, Department of Electrical Engineering, Faculty of Engineering, J.N.V. University, Jodhpur-342 011 India (email: jvajpai@gmail.com).

which attempts to reduce the error. This process is repeated through a series of successive trees. The final predicted value is formed by adding the weighted contribution of each tree.

The Tree Boost algorithm generates the most accurate models with minimum over fitting if only a portion of the data rows are used to build each tree in the series. This is the stochastic part of stochastic gradient boosting. Usually, the individual trees are fairly small (typically 3 levels deep with 8 terminal nodes), but the full Tree Boost additive series may consist of hundreds of these small trees.

IV. PROBLEM STATEMENT

DAMADICS (Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems) benchmark has been developed as a benchmarking tool for fault diagnosis and isolation (FDI) methods . The core of this benchmark is a Simulink model of an electro-pneumatic valve actuator. This model includes three subsystems: a control valve, a spring-and-diaphragm pneumatic servomotor, and a positioner. The servomotor acts on the control valve plug which position controls the fluid flow passing through the pipelines. The stem of the servomotor is driven by compressed air, which acts on a flexible diaphragm and is balanced by a spring. A positioner is used to avoid miss-positions of the stem caused by internal and external factors like friction and change of supply pressure and provides digital I/O for the actuator.

The benchmark contains total 44 types of fault scenarios, but as reported in the literature [10], the misclassification occurs due to overlapping phenomenon among different fault classes.

Koscielny et al [11] have carried out study of Fault detectability and distinguishability for DAMADICS Benchmark Actuator. The results obtained by them show that fault distinguishability of actuator can be improved due to the application of the three-valued residual evaluation instead of a binary one.

In this paper, after using FIS approach the authors have categorized following five sets of faults as unconditionally indistinguishable faults:

{F 2; F 5; F 7}; {F 3; F 6; F 18}; {F 4; F 8}, {F 9; F 10}; {F 13; F 15}.

The work presented here is a sincere attempt for further improvement of fault diagnosis results obtained in the cited work on DAMADICS benchmark. An attempt has been made to further isolate the above five sets of faults using tree boost technique.

V. METHODOLOGY

The dataset used for this case study have been generated by employing the MATLAB-SIMULINK model of the actuator as shown in fig 2.

In accordance with the scope of the defined objective for this paper, only data related fault categories F2(valve or valve seat sedimentation),F3(valve or valve seat erosion),F4(increase of valve friction),F5 (External leakage: leaky bushing, covers, terminals), F6(internal leakage),F7(medium evaporation or critical flow),F8 (Twisted piston rod),F9 (servomotor housing or terminal tightness), F10(servomotor diaphragm perforation), F13(stem displacement sensor fault), F15 (positioner spring fault) & F18(fully or partly opened bypass valves) have been considered.

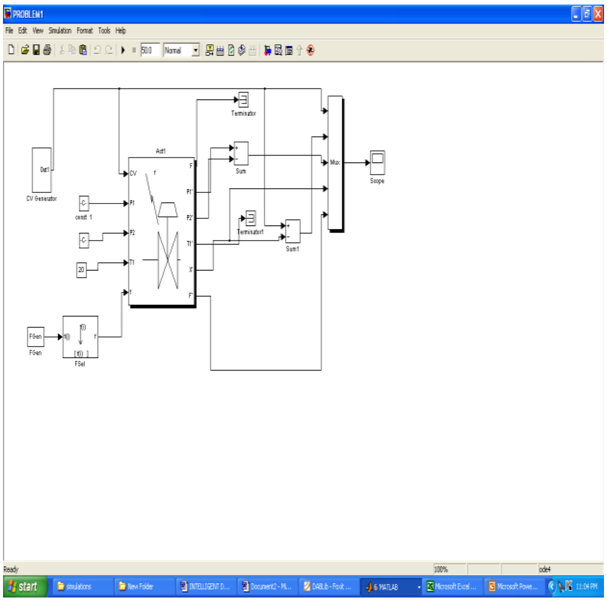

Fig2: MATLAB- Simulink Model

The tree boost series model [12] has been used for this purpose, with Maximum splitting levels of 5 and maximum trees in tree boost series limited to 400. Classification analysis has been performed while using surrogate splitters for any missing values in dataset. The category weights or priors were obtained from data file distribution and variable weights were set to be initially equal. Misclassification cost was also set to be equal or unitary. Random sampling (20%) was used for validation. The tree pruning criterion was selected to be minimum absolute error. Summary of variables have been presented in Table1.

[Table 1: Summary of Variables]

No.	Variable	Class	Type
1	CV	Predictor	Continuous
2	P1	Predictor	Continuous
3	P2	Predictor	Continuous
4	T	Predictor	Continuous
5	X	Predictor	Continuous
6	F	Predictor	Continuous
7	Type of fault	Target	Categorical

VI. RESULTS

The results obtained using Stochastic Gradient Boosted Decision Trees based method for the five sets of unconditionally indistinguishable faults have been presented in Tables 2-6 along with brief account of Tree boost model summary for each case.

(1) Model Summary for Fault set:-{F2; F5; F7}

[Table 2: Misclassification Table]

Actual			Misclassified		
Category	Count	Wt.	Count	Wt.	%
F2	16	16	0	0	0
F5	16	16	0	0	0
F7	16	16	0	0	0
Total	48	48	0	0	0

The minimum error occurs with 11 trees.
The minimum point is smoothed by 5 trees.
The specified minimum number of trees is 10.
The tree series will be pruned to 12 trees.
Average number of group splits in each tree = 11.5

(2) Model Summary for Fault set :-{F3; F6; F18}

[Table 3: Misclassification Table]

Actual			Misclassified		
Category	Count	Wt.	Count	Wt.	%
F3	16	16	3	3	18.75
F6	16	16	3	3	18.75
F18	16	16	3	3	18.75
Total	48	48	9	9	18.75

The minimum error occurs with 139 trees.
The minimum point is smoothed by 5 trees.
The specified minimum number of trees is 10.
The tree series will be pruned to 17 trees.
Average number of group splits in each tree = 11.3

(3) Model Summary for Fault set: -{F 4; F 8}

[Table 4: Misclassification Table]

Actual			Misclassified		
Category	Count	Wt.	Count	Wt.	%
F4	16	16	4	4	25.00
F8	16	16	5	5	31.25
Total	32	32	9	9	28.125

The minimum error occurs with 398 trees.
The minimum point is smoothed by 5 trees.
The specified minimum number of trees is 10.
The tree series will be pruned to 25 trees.
Average number of group splits in each tree = 2.4

(4) Model Summary for Fault set:-{F9; F 10}

[Table 5: Misclassification Table]

Actual			Misclassified		
Category	Count	Wt.	Count	Wt.	%
F9	16	16	0	0	0.00
F10	16	16	0	0	0.00
Total	32	32	0	0	0.00

The minimum error occurs with 54 trees.
The minimum point is smoothed by 5 trees.
The specified minimum number of trees is 10.
The tree series will be pruned to 70 trees.
Average number of group splits in each tree = 2.3

(5) Model Summary for Fault set:-{F13; F 15}

[Table 6: Misclassification Table]

Actual			Misclassified		
Category	Count	Wt.	Count	Wt.	%
F13	16	16	0	0	0.00
F15	16	16	1	1	6.25
Total	32	32	7	7	3.125

The minimum error occurs with 143 trees.
The minimum point is smoothed by 5 trees.
The specified minimum number of trees is 10.
The tree series will be pruned to 11 trees.
Average number of group splits in each tree = 2.3

VII. DISCUSSION

Tree Boost models often can provide greater predictive accuracy than single-tree models, but they have the disadvantage that they cannot be visualized like a single tree i.e. Tree Boost models are more like a black box. Because of this, it is advisable to create both a single-tree and a Tree Boost model. The single-tree model can be studied to get an intuitive understanding of how the predictor variables relate, and the Tree Boost model can be used to score the data and generate highly accurate predictions.

The Tree Boost algorithm is functionally similar to decision tree forests because it creates a tree ensemble, but a Tree Boost model consists of a series of trees whereas a decision tree forest consists of a collection of trees grown in parallel. Tree Boost generates a series of trees with the output of one tree going into the next tree in the series. In contrast, a decision tree forest grows a number of independent trees in parallel, and they do not interact until after all of them have been built. Both Tree Boost and decision tree forests produce high accuracy models. Experiments have shown that Tree Boost works better with some applications and decision tree forests with others, so it is best to try both methods and compare the results.

Hence, results have been obtained for above considered five fault sets using Single decision tree, Decision tree forests and Tree Boost approach and a comparative statement of results have been shown in table7. Also, a single decision tree model for First fault set {F2, F5, and F7} has been shown in fig 3.

[Table 7: Comparison with Decision Tree and Tree forest approach]

Fault set	% Miscla-ssified Using Single Decision Tree Approach	% Miscla-ssified Using Tree Forest Approach	% Miscla-ssified Using Tree Boost Approach
{F2,F5,F7}	3.33 %	6.67%	0%
{F3,F6,F18}	30 %	68.33%	18.75%
{F4,F8}	45%	100%	28.12%
{F9,F10}	35%	52.5%	0%
{F13,F15}	0%	12.5%	3.12%

The strength of proposed Tree Boost Approach lies in the accuracy manifested in handling the classification (discrimination) task for unconditionally indistinguishable fault sets with fine precision.

Future research needs to focus on further improvement of fault diagnosis results on DAMADICS benchmark. One possible direction in which authors are presently working is to investigate the improvement in performance of the fault diagnosis task using perception based decision making.

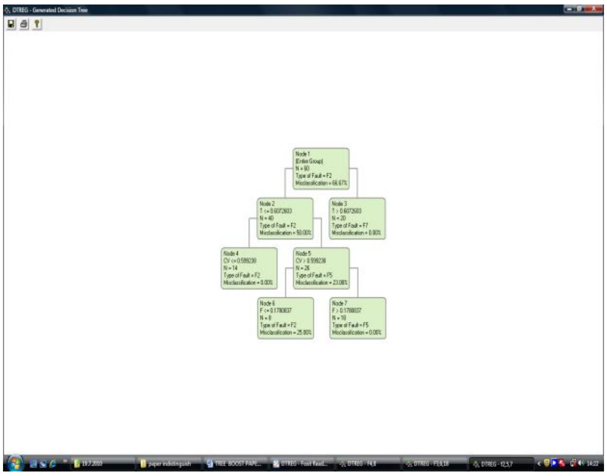

Fig 3: Single Decision Tree Model for {F2, F5, F7}

References

[1] Friedman, Jerome H., “Greedy Function Approximation: A Gradient Boosting Machine” Technical report, Dept. of Statistics, Stanford University. 1999.

[2] S.M.Weiss and C.A. Kulikowski, “Computer Systems that Learn: Classification and Prediction Methods from Statistics”, Neural Nets, Machine Learning, and Expert Systems, 1991.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J.Stone. “Classification and Regression Trees”. Pacific Grove: Wadsworth, 1984.

[4] M. Mehta, R. Agrawal, and J. Rissanen.,“SLIQ: A fast scalable classifier for data mining” In Proc. Of EDBT, 1996.

[5] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, “An empirical comparison of decision trees and other classification methods” TR 979, Department of Statistics, UW Madison, June 1997.

[6] Quinlan, J. R., “Induction of decision trees. In Machine Learning” pp. 81–106, 1986.

[7] Chen,K.,LU,R.,Wong,C.K.,Sun,G.,Heck,L., and Tseng,B.L. Trada, “Tree based ranking function adaptation”. In CIKM, pp. 1143– –1152, 2008

[8] Zheng,Z.,Chen,K.,Sun,G., and Zha, H. “A regression framework for learning ranking functions using relative relevance judgments”. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval , pp. 287–294,2007.

[9] Friedman, J. H., “Stochastic gradient boosting”. Comput. Stat. Data Anal. 38,4 , pp. 367–378, 2002.

[10] Michal Bartys et al, “Introduction to the DAMADICS actuator FDI benchmark study” Control Engineering Practice 14 , pp. 577–596, 2006

[11] Jan M. Koscielnya, Micha Bartys, Pawe Rzepiejewski Jose Sa da Costa, “Actuator fault distinguishability study for the DAMADICS benchmark problem “, Control Engineering Practice 14 , pp. 645-652, 2006

[12] Phillip H. Sherrod, “DTREG Predictive Modeling Software”, 2003.