Quick and Dirty Quantum Predictions with FIRMplus

Christophe G. Lambert
Golden Helix Inc.
P. O. Box 10633
Bozeman, MT 59718
lambert@goldenhelix.com

James E. Grover
Golden Helix Inc
P. O. Box 10633
Bozeman MT 59718
grover@goldenhelix.com

ABSTRACT

In this paper, we describe our approach to the Quantum Physics classification problem posed in KDD Cup 2004. We used our FIRMplus recursive partitioning random tree approach to classify the two particle types. We spent almost no time tuning parameters, which made our 3rd place finish somewhat surprising.

Keywords

FIRM, random FIRM, recursive partitioning.

1. INTRODUCTION

At the heart of the subject specific data analysis tools our company has developed is a computational engine based on the Formal Inference-based Recursive Modeling (FIRM) recursive partitioning methodology [2]. In our extended version of FIRM, FIRMplus™ , we have included a very fast segmenting algorithm, as well as a multiple random tree prediction capability. Successful applications of FIRM abound in the sciences in such diverse areas as archaeology, cheminformatics, environmetrics, geology, bioinformatics, sociology, and zoology [6,9-11]. We were pleasantly surprised at our prediction performance for this physics task considering our lack of subject matter knowledge, and how little we customized our approach to this problem.

2. BACKGROUND

Early work on FIRM dates back to the 1970's where the algorithms were laid out by Hawkins [5] on how to optimally segment data. FIRM was released in 1982, with significant advances made by Hawkins and Musser over the years [3,4,7,8].

With a dependent *n*-vector *y*, and a covariate vector *x*, there are $\binom{n}{k-1}$ ways to segment the elements of *y* into *k* bins so as to minimize the sum of squared deviations from the mean within each segment. The FIRM approach uses dynamic programming to find the optimal *k*-way segmenting, splits the data into the *k* subsets, and then recursively segments each subset until a stopping criterion is reached. The stopping criterion is based on a hypothesis test that determines the probability that the means or proportions of the dependent vector are statistically equivalent between segments.

We have extended the original FIRM approach to use the appropriate binary log-likelihood model [4]. We used the NextFIRM multiple tree prediction approach developed by Musser [3,8], and we have devised a proprietary probabilistic dynamic programming approach to reduce the $O(kn^2)$ running time of the Hawkins segmenting algorithm to $O(kn^{1.5})$.

Some characteristics of the FIRM approach that differentiate it from other recursive partitioning approaches such as CART [1] are the ability to find optimal multiway splits for continuous or ordinal data, a method for handling "predictive missingness" via a floating category for missing data, as well as its formal statistical hypothesis testing approach to model building.

3. APPROACH

There was little chance of overfitting with our approach. We performed a single experiment to determine if binary splits gave better predictions than 3-way splits. We used 50% of the training data (25,000 observations) to build a model, and used the other 50% to validate the model.

We used the 78 provided covariates with their associated missing values in the model building, and generated two sets of 100 random trees, one with 2-way splits and one with 3-way splits. The random tree generation approach is to select at random from among the top 10 statistically significant variables at each node in the tree and split using that variable. Recursively within each daughter node, the best splits are found and one is chosen at random, and the process continues until there are no statistically significant splits to be performed at the leaf nodes. Since we have found in the past that overbuilding the trees gives the best prediction results, we used a lax Bonferroni-adjusted p-value threshold of $p<0.99$ as a stopping criteria for tree-building.

A prediction for a tree is the mean of the observations of the leaf node in which an observation falls. To get a multiple tree prediction for a single observation, we use the covariates of the observation to drop it down each tree and find the leaf node in which it falls. We then average the means of those leaf nodes from the training set to get the prediction for the test observation. Because the response in this case is binary, the resulting prediction ranges from [0...1]. We did no post-processing of results.

4. RESULTS

We found that 3-way splits outperformed 2-way splits, categorizing about 0.5% more observations correctly in the test set. We then used the entire set of 50,000 training examples and built 40 random trees with 3-way splits and applied the resulting trees to predict the 100,000 unknown contest observations. We submitted the same predictions for all 4 scoring criteria.

The prediction accuracy we attained within our 50-50 train-test sampling of the training set was no different than the results we achieved when extrapolating our model to the final test set of 100,000 particles, about 72.8%.

5. CONCLUSION

Could we have done better if we built more trees, created additional covariates that were sums or products of the original 78, or tuned our parameters some more? Perhaps, but the bottom line is that with a quick and dirty default set of parameters, our results ranked with the best prediction attempts even though we knew virtually nothing about the problem domain and spent only a couple of hours on the problem. The winners are to be commended for squeezing out the last epsilon in prediction possibility this data afforded. We nevertheless feel a sense of accomplishment in earning 3rd place with an off-the-shelf solution.

6. REFERENCES

[1] Breiman, L.; Friedman, J.H.; Olshen, R. A.; Stone, C. J. (1984) "Classification and Regression Trees", Wadsworth International, Belmont.

[2] Hawkins, D. M. and G. V. Kass (1982). "Automatic interaction detection". In D. M. Hawkins (Ed.), Topics in Applied Multivariate Analysis. Cambridge University Press.

[3] Hawkins, D. M. and Musser, B. J., (1999), "One tree or a forest? Alternative dendrographic models", Computing Science and Statistics, 30, 534--542.

[4] Hawkins, D. M. (2001) "Fitting multiple change-points to data", Computational Statistics and Data Analysis, 37, 323-341.

[5] Hawkins, D.M. (1972) "On the choice of segments in piecewise approximation", J. Inst. Maths Applics, v. 9, 250-256.

[6] Hawkins, D.M. (2003) "Divide and model", Scientific Computing World, 70 68-69.

[7] Hawkins. D. M., and Musser, B. J. (2001), "Feature selection with nondeterministic recursive partitioning", Proceedings of the American Statistical Association [CD-ROM] Alexandria, VA: ASA.

[8] Musser, B. J. (1999) "Extensions to Recursive Partitioning", Ph.D. Thesis, University of Minnesota School of Statistics.

[9] Rusinko, A. III; Farmen, M.W.; Lambert, C.G.; Brown, P.L. and Young, S.S. (1999), Journal of Chemical Information and Computer Sciences v. 39 no. 6 pp. 1017-1026.

[10] Young, S.S.; Ekins, S. and Lambert, C.G. (2002) "So many targets so many compounds, but so few resources", Current Drug Discovery, December, 17-22.

[11] Young, S.S.; Gombar, V.K.; Emptage, M.R.; Cariello, N.F. and Lambert, C.G. (2002) "Mixture De-Convolution and Analysis of Ames Mutagenicity Data", Chemometrics and Intelligent Laboratory Systems, v.60 pp. 5-11.

About the authors:

Christophe G. Lambert is the President and CEO of Golden Helix Inc. He earned his Ph.D. in computer science from Duke University in 1997.

James E. Grover is a Scientific Applications Programmer at Golden Helix. He earned his M. S. in Applied Mathematics from the California Institute of Technology in 1972.