Evolutionary Linear Discriminant Analysis for Multiclass Classification Problems

Michael F. Korns
Lantern Credit LLC
2240 Village Walk Drive Suite 2305
Henderson, Nevada 89052
mkorns@korns.com

ABSTRACT

This paper implements Linear Discriminant Analysis (LDA) together with genetic programming symbolic classification for financial multiclass classification problems. A very brief description of the matrix theory of LDA is included. The implementation details in an industrial strength multiclass classification system are discussed. The algorithm is tested using statistically correct, out of sample training and testing. The algorithm's behavior is demonstrated on real world problems previously published as UCI test suites and financial real world problems.

KEYWORDS

Symbolic Regression and Classification, Genetic Programming, Linear Discriminant Analysis.

1 INTRODUCTION

An effort is underway to extend the reach of Symbolic Regression (SR) [3] into the realm of multiclass classification. In this paper we propose that standard Linear Discriminant Analysis (LDA) [11], [12], [13] be seriously considered as one of the GP assisted fitness training techniques used in future Symbolic Classification systems (SC). LDA is widely used in academia and industry and would be readily accepted as a GP assisted fitness training technique in future symbolic classification systems.
In the past few years, symbolic regression has advanced into the early stages of commercial exploitation with impressive accuracy results [2], [7], [8]. Our use of SR in investment finance has been particularly rewarding. However, extending SR to perform multi-class classification has been problematic [1], [10]. Most industrial strength SR systems use simple and multiple regression with general linear models as GP assisted training algorithms [4], [5], [6]. Unfortunately neither simple regression nor multiple regression perform at all well in multiclass classification problems.

GECCO '17 Companion, July 15-19, 2017, Berlin, Germany
ACM 978-1-4503-4939-0/17/07.
http://dx.doi.org/10.1145/3067695.3075964

Several researchers have developed the M2GP algorithm for GP assisted fitness training in multiclass classification [1] and the M3GP algorithm [10]. Other research has developed the evolutionary MDC algorithm for GP assisted fitness training in multiclass classification [9].
In this paper we implement Linear Discriminant Analysis (LDA) as a GP assisted fitness training algorithm for multiclass classification problems. Linear Discriminant Analysis assisted Symbolic Classification performs well on a number of test problems including interesting financial problems.

2 BRIEF LDA BACKGROUND

Linear Discriminant Analysis (LDA) is a generalization of Fischer's linear discriminant, which is a method to find a linear combination of features which best separates K classes of training points [11], [12], [13]. LDA is used extensively in Statistics, Machine Learning, and Pattern Recognition. We use Bayes rule to minimize the classification error percent (CEP), which is defined as the count of erroneous classifications divided by the size of Y, by assigning a training point $X_{[n]}$ to the class k if the probability of $X_{[n]}$ belonging to class k, $P(k|X_{[n]})$, is higher than the probability for all other classes as follows.

- $EY[n] = k$, iff $P(k|X_{[n]}) \geq P(j|X_{[n]})$ for all $1 \leq j \leq K$

The CEP is computed as follows.

- $CEP = \sum(EY_{[n]} \neq Y_{[n]} \mid \text{for all } n) / N$

Therefore, each discriminant function D_k acts a Bayesian estimated percent probability of class membership in the formula.

- $y = \text{argmax}(D_1, D_2, \ldots, D_K)$

The technique of LDA makes three assumptions, (a) that each class has multivariate Normal distribution, (b) that each class covariance is equal, and (c) that the class covariance matrix is nonsingular. Once these assumptions are made, the mathematical formula for the optimal Bayesian discriminant function is as follows.

- $D_k(X_n) = \boldsymbol{\mu}_k (C_k)^{-1} (X_n)^T - 0.5 \boldsymbol{\mu}_k (C_k)^{-1} (\boldsymbol{\mu}_k)^T + \ln(P_k)$

Where X_n is the nth training point, $\boldsymbol{\mu}_k$ is the mean vector for the kth class, $(C_k)^{-1}$ is inverse of the covariance matrix for the kth class, $(X_n)^T$ is the transpose of the nth training point, $(\boldsymbol{\mu}_k)^T$ is the transpose of the mean vector for kth class, and $\ln(P_k)$ is the natural logarithm of the naïve probability that any training point will belong to the kth class. A central aspect of LDA is that each discriminant function is a linear variation of every other discriminant function, with B basis functions, such that each discriminant function has the following format.

- $D_k = c_{k0} + c_{k1} * Bf_1 + c_{k2} * Bf_2 + \ldots + c_{kB} * Bf_B$

The K*(B+1) coefficients are selected so that the ith discriminant function has the highest value when the y = i (i.e. the class is i). LDA is the mathematical technique used for selecting the optimized coefficients c_{00} to c_{KB}.

3 TEST PROBLEMS

In this section we apply the LDA algorithm on test problems downloaded from the University of California at Irvine machine learning repository https://archive.ics.uci.edu/ml/datasets.html. and downloaded from the Broad Institute cancer data sets http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Another Volatility data set was constructed from the Yahoo down loadable VIX and UVXY daily historical data sets. This test problem attempted to classify the next day's profit or loss in the UVXY ETF, entirely from the previous day's percent change in the VIX and the percent change in the 140 day moving average of the VIX.

Test Problems

- (T1) Acute Myeloid Leukemia (from Broad Institute)
- (T2) Iris (from UCI)
- (T3) Heart Disease (from UCI)
- (T4) Volatility (from Yahoo VIX & UVXY data)
- (T5) Bank Marketing (from UCI)

Table 1: Real Data Test Problem Results

Test	WFFs	Train-Hrs	Train-CEP	Test-CEP
T1	421K	3.01	0.0000	0.0183
T2	223K	1.30	0.0000	0.0000
T3	471K	3.56	0.0815	0.1051
T4	341K	2.14	0.0472	0.0628
T5	258K	1.94	0.1263	0.0972

Notes. The number of regression candidates tested before finding a solution is listed in the Well Formed Formulas (**WFFs**) column. The elapsed hours spent training on the training data is listed in the (**Train-Hrs**) column. The classification error percent fitness score of the champion on the training data is listed in the (**Train-CEP**) column. The classification error percent fitness score of the champion on the testing data is listed in the (**Test-CEP**) column with **0.0593** average testing fitness.

By enhancing the baseline Pareto front GP algorithm with LDA assisted fitness training, we achieve attractive classification error scores on all test problems. Furthermore, the Volatility test data achieved very good CEP training and testing scores, and categorized, without any losses, the day's when the UVXY next day profit was 15% or above. There were fourteen estimated trading signals all of which resulted in next day UVXY profits of 15% or more.

4 CONCLUSIONS

Several papers have proposed GP assisted fitness training techniques for multiclass classification problems [1], [9], [10]. In this paper, Linear Discriminant Analysis is proposed as a GP assisted fitness training technique [13]. Testing the LDA algorithm on several test problems resulted in quite reasonable classification scores. Results on investing finance problems were particularly promising.

The next steps should be the comparison of each of the proposed GP assisted fitness training techniques (M2GP, M3GP, MDC, and LDA) on real world test problems with the goal of creating an industrial strength future symbolic classification system for multiclass classification problems. Once we have an understanding of the behavior of each of these GP assisted fitness training techniques on different classes of problems, we can then compare the classification behavior of GP assisted multiclass classification against the current leading industry classification techniques (Neural Nets, Decision Forests, Deep Learning, etc.).

REFERENCES

[1] Ingalalli, Vijay, Silva, Sara, Castelli, Mauro, Vanneschi, Leonardo 2014. A Multi-dimensional Genetic Programming Approach for Multi-class Classification Problems. *Euro GP 2014* Springer.

[2] Korns, Michael F. 2013. Extreme Accuracy in Symbolic Regression. *Genetic Programming Theory and Practice XI*. Springer, New York, NY..

[3] Koza, John R. 1992. *Genetic Programming: On the Programming of Computers by means of Natural Selection*. The MIT Press. Cambridge, Massachusetts.

[4] Korns, Michael F. 2012. A Baseline Symbolic Regression Algorithm. *Genetic Programming Theory and Practice XI*. Springer, New York, NY.

[5] Keijzer, Maarten. 2003. Improving Symbolic Regression with Interval Arithmetic and Linear Scaling. *European Conference on Genetic Programming*.

[6] Billard, Billard., Diday, Edwin. 2003. *Symbolic Regression Analysis*. Springer. New York, NY..

[7] Korns, Michael F. 2014. Extremely Accurate Symbolic Regression for Large Feature Problems. *Genetic Programming Theory and Practice XII*. Springer, New York, NY.

[8] Korns, Michael F. 2015. Highly Accurate Symbolic Regression for Noisy Training Data. *Genetic Programming Theory and Practice XIII*. Springer, New York, NY.

[9] Korns, Michael F. 2016. An Evolutionary Algorithm for Big Data Multiclass Classification Problems. *Genetic Programming Theory and Practice XIV*. Springer, New York, NY.

[10] Munoz, Louis, Silva, Sara, M. Castelli, Trujillo 2014. M3GP Multiclass Classification with GP. *Euro GP 2015* Springer.

[11] Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics 7 (2) 179-188*.

[12] Friedman, J. H. 1989. Regularized Discriminant Analysis. *Journal of American Statistical Association 84 (405) 165-175*.

[13] McLachan, Geoffrey, J. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley. New York, NY.