A Machine Learning-Based Assessment of the Number of Contributors in DNA Mixtures

Michael A. Marciano, M.S. [1,*], Jonathan D. Adelman, M.S.[1,*]

[1]Forensic & National Security Sciences Institute, College of Arts and Sciences, Syracuse University, 107 College Place 1-014 CST, Syracuse, NY, 13244, USA

*Correspondence: mamarcia@syr.edu (M.A. Marciano) and jdadelma@syr.edu (J.D. Adelman).

Introduction

An accurate estimation of the number of contributors in a DNA mixture is central to the downstream deconvolution of the sample. Incorrect NOC estimations can negatively impact both likelihood ratios and the ultimate sample deconvolution [1, 2]. Traditionally, the maximum allele count method has been used to predict the number of contributors however, this method loses reliability when the samples have three or greater contributors [3-7]. Maximum likelihood estimation [8] and Markov chain Monte Carlo [9] methods have also been proposed. More recently, a machine learning based method for the estimation of the number of contributors was proposed [10].

Machine learning - a branch of artificial intelligence - is the systematic study of algorithms and systems that improve their knowledge or performance with experience [11]. A machine learning algorithm can, after exposure to an initial set of data, be used to generalize; meaning that it can create a predictive model capable of evaluating new, previously unseen examples. Machine learning is a widely-used approach with an incredibly diverse range of applications, including object recognition [12], natural language processing [13], and DNA sequence classification [14]. It is ideally suited for classification problems involving implicit patterns, and is most effective when used in conjunction with large amounts of data. Although machine learning has not previously been used within the domain of DNA mixture analysis, the problem area of determining the number of donors in a DNA mixture is well-suited to such an endeavor due to two key problem characteristics: there exists a large repository of human DNA mixture data in electronic format, and these data are high-dimensional and complex; patterns in such data are often non-obvious and beyond the effective reach of manual analysis but can be statistically evaluated using a machine learning algorithm.

The goal of machine learning is to find the hypersurface - or for simpler, two-dimensional problems in feature space, the decision boundary - that best separates classes of samples. PACE, for example, separates single-source samples from two-person mixtures, and those two classes are separated from three-person mixtures, etc. We can use Fisher's canonical Iris data set as context for evaluating this goal. When there are only two classes (in this case, Iris species) to separate and two features (in this case, sepal width and petal height) used to separate them by, one can view the classification problem using a traditional scatterplot (Figure 1). That two-dimensional plot in feature space becomes a cube when we add an additional feature (Figure 2), however, and high-dimensional feature space simply cannot be successfully navigated by a human analyst. This is where machine learning algorithms excel; they find the optimal hypersurface in high-dimensional feature space.

Within the specific context of estimating the number of contributors, such an approach not only minimizes prediction error but also produces mathematically valid probability approximations [15]. This approach is fully continuous; peak height ratios, for example, are an integral part of the initial construction of a learning algorithm's feature vector. The approach also leverages the general characteristic among machine learning algorithms of being able to find implicit patterns in complex data. And not least, such an approach is very fast, often taking only seconds to evaluate complex DNA mixtures. The one major drawback to such an approach is its reliance on a massive training data set used to learn the optimal model for the classification problem. This data library must be comprehensive in its portrayal of DNA mixtures, lest the eventual predictive model fare poorly when classifying mixtures substantively different from those it had initially been exposed to. Such a requirement in PACE's case has led to the collection of thousands of DNA samples.

Methods

It is inefficient for a machine learning to analyze raw data in many cases. With a fixed number of training samples, the predictive power of the algorithm reduces as the problem's dimensionality increases; this is known as the Hughes phenomenon [16], or as Bellman "curse of dimensionality" [17]. As a practical effect, the continual addition of new features to a feature vector ultimately leads to decreased accuracy of the resulting classification model; one cannot use a brute force solution and simply use the entirety of a complex data set for machine learning. The solution usually relied upon by data scientists is to intelligently reduce the dimensionality of the problem by compressing the data into a vector of relatively few, high-information features. In PACE's case these features are both native to the underlying data set (e.g. locus-specific and sample-wide peak counts) and derived from manipulations of the data set.

Several algorithms have been successfully used in conjunction with PACE, two of which are the multilayer perceptron [18] and the support vector machine [19]. The Multilayer Perceptron (MLP) algorithm is an artificial neural network in which backward propagation of errors is used to train the network’s weights and thresholds. In this study, a single hidden layer of neurons was used, and the four output nodes correspond to the four classes of number of contributors. The Support Vector Machine (SVM) algorithm attempts to optimize classification by maximizing the distance between the margins of classes. There are both linear and non-linear versions of this classifier, the latter of which is specifically designed for classes, such as the number of contributors, that cannot be linearly separated.

A machine learning algorithm will “learn” a predictive model, and that model in turn, is potentially capable of classifying new, unfamiliar data. This ability to predict outcome values for previously unseen data is termed generalization. Merely providing training data to a learning algorithm is an insufficient generalization strategy; the algorithm may end up learning specific patterns only found in the training data by chance, and would then erroneously leverage those patterns to aid with classification. To ensure generalization and avoid potential overfitting, the learning algorithm must be both trained and its predictive model must be tested, and it is the resulting testing accuracy, not the training accuracy, that serves to validate the learned model. Such an approach requires that the initial data set (in this case, the library of DNA mixtures) must correspondingly be partitioned into completely separate training and testing subsets. For all modeling efforts herein, the training data set was created by randomly selecting 75% of the initial DNA samples, with the other 25% used for testing how generalizable the learned model is.

Machine learning algorithms contain hyperparameters which can be loosely thought of as knobs that tune the algorithm and thereby affect its behavior. Some hyperparameters can have a nontrivial impact on the resulting training time or even classification accuracy. Any attempt to tune these hyperparameters and thereby maximize an algorithm's classification accuracy are typically accompanied by a further partitioning of training data to ensure that data used for "tuned" algorithm validation are not also used to validate the final model. A viable alternative to data partitioning is the technique of k-fold cross-validation. In this technique the algorithm is trained a total of k times, with a fraction 1/k of training examples left out each time for validation purposes [20], leading to k distinct "folds". Each fold provides summary metrics that describe the algorithm's performance for that particular training, and the results from each of the folds are averaged to provide an overall assessment of model performance. All hyperparameter tuning in this study utilizes 5-fold cross-validation on the training data set.

In classification problems, the training data contain unequal instances for different classes. Four-contributor samples, for example, are less common than three-contributor samples, which are less common than two-contributor samples. Machine learning algorithms are often sensitive to imbalance in the predictor classes, and will bias the prediction model towards the more common class. One solution is to perform oversampling, whereby an oversampling algorithm generate additional samples for the less-populous classes based on the characteristics of existing data. PACE uses random oversampling for multi-class sampling and SMOTE for binary oversampling [21].

Materials

Sample sets included 1035 samples amplified with PowerPlex Fusion amplification kit (Promega Corp) and 1405 samples amplified with the Identifiler® amplification kit (Thermo Fisher Scientific). Additional details can be found in table 1.

Table 1: Sample sets used for PACE analyses.

	Identifiler®	**PowerPlex® Fusion**
Sample #	1405	1035
Individuals	16	45
Template Range	12.5pg – 10.0ng	3.25pg – 4.0ng
Mixture Ratios	28	38
Instruments	5 (3100 & 3130)	6 (3100,3130 & 3500)
Injection time / voltage	6 times / 2 kVs	2 times / 5 kVs

Each sample is analyzed by PACE and probabilities associated with each class is returned, i.e. a sample will have a probability associated with NOC 1, 2, 3 and 4.

Results

Processing speed

Five samples were run using a64-bit Windows 10 laptop computer with an Intel Core i7 2.70gHz processor and 16GB of RAM. The four contributor samples with 6.0ng of template DNA amplified using the Identifiler amplification kit had the longest processing time (2 minutes 8s) due to the complexity of the resulting electropherogram (Table 2).

Table 2: PACE processing speed.

Sample	Expected NOC	Expected Ratio	ng amplified	Run time
4_BU_0.2	4	1.6 : 3 : 1 : 1	0.2	37.8s
4_BU_6	4	1.6 : 3 : 2 : 2	6.0	2m 8s
280	3	1 : 1 : 3	0.15	7.3s
981	3	6 : 1 : 1	2.0	3.2s
501	2	19 : 1	0.0625	2.8s

Number of contributor estimation

Tables 3 and 4 include the results for the performance of PACE for Identifiler and preliminary results for PowerPlex Fusion, respectively. All “correct” calls are based on the maximum probability of NOC, i.e. the class with the highest probability returned by PACE. There were no over/under estimations by greater than one contributor.

Table 3: PACE- Identifiler results.

		***PACE* Predicted** **Number of Contributors**				**Percent Correct**
		1	**2**	**3**	**4**	
Expected **Number of Contributors**	**1**	***94***	*0*	*0*	*0*	***100%***
	2	***3***	***152***	*0*	*0*	***98.1%***
	3	*0*	***2***	***72***	*0*	***97.3%***
	4	*0*	*0*	*0*	***29***	***100%***

Table 4: PACE- PowerPlex Fusion results.

		PACE Predicted **Number of Contributors**				**Percent Correct**
		1	**2**	**3**	**4**	
Expected **Number of Contributors**	**1**	***68***	*0*	*0*	*0*	***100 %***
	2	***0***	***107***	***1***	*0*	***99 %***
	3	*0*	***4***	***58***	*2*	***91 %***
	4	*0*	*0*	***2***	***28***	***93 %***

Comparison to Maximum Allele Count Methods

NOC accuracy was compared between PACE-Identifiler (Figure 3), PACE Fusion (Figure 4) and maximum allele count methods using the following thresholds: 50rfu, 100 rfu, and 150rfu and a dynamic threshold calculated based on sample-locus noise.

Conclusions

The machine learning approach represents a fully continuous and probabilistic approach to the prediction of the number of contributors. The method is highly accurate, with over 90% accuracy when predicting three and 4+ contributor samples. Computational resources will not need to change; existing computing resources will be sufficient to run PACE and will lead to rapid results in approximately 10-15seconds per sample.

Acknowledgements

We would like to thank the New York City Office of the Chief Medical Examiner Department of Forensic Biology, the Palm Beach Sheriff's Office, Kentucky State Police, the Promega Corporation, Michigan State Police, the Onondaga County Center for Forensic Sciences, the National Institute of Justice and Niche Vision LLC. We would also like to thank Victoria Williamson and Laura Haarer.

Figures

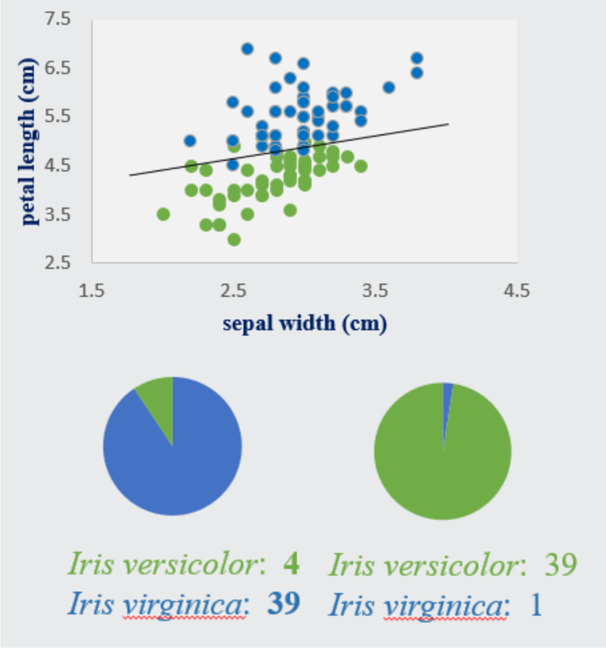

Figure 1: A scatterplot showing the relationship between sepal width and petal height for two Iris species.

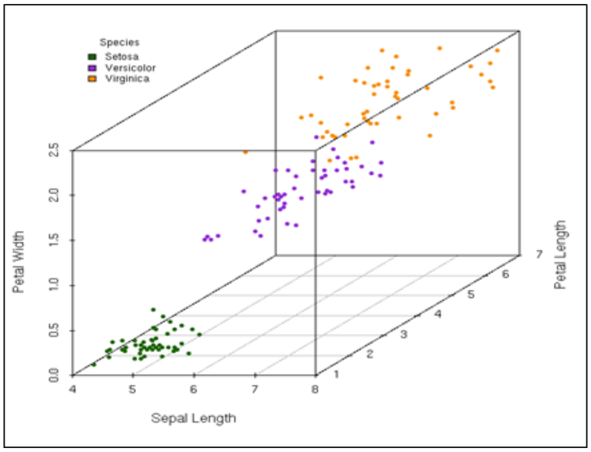

Figure 2: A cube showing the relationship between sepal width, petal width, and petal length for three Iris species.

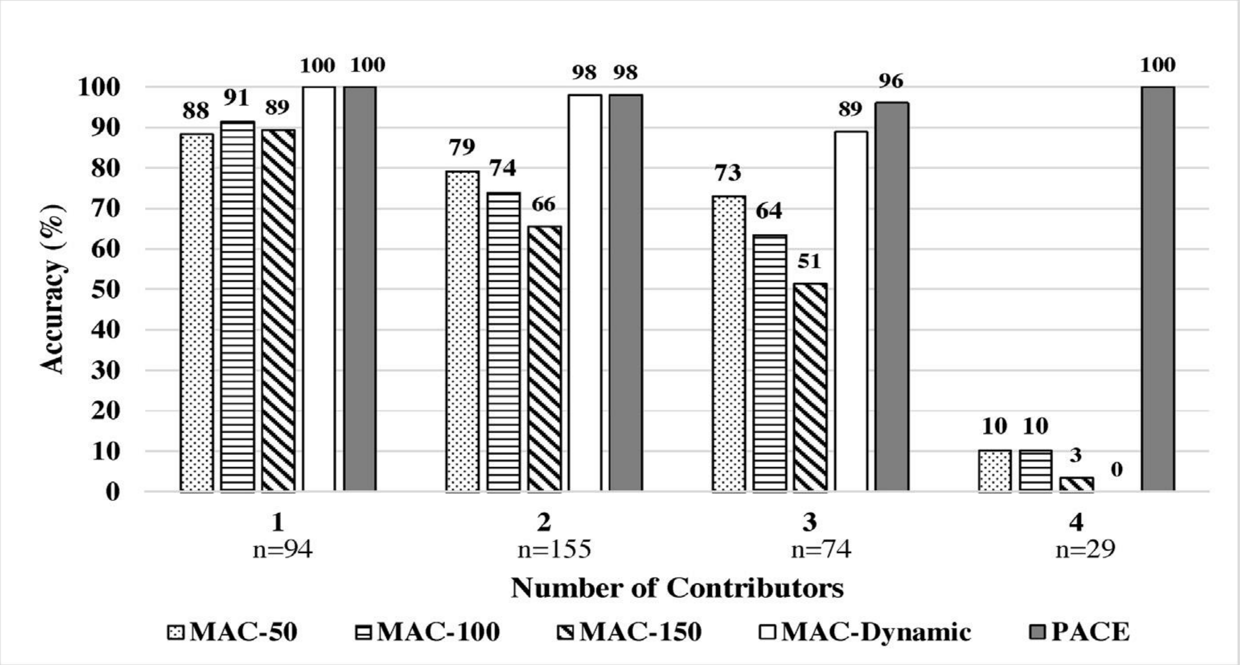

Figure 3: Comparison of PAVE-Identifiler to MAC. [10]

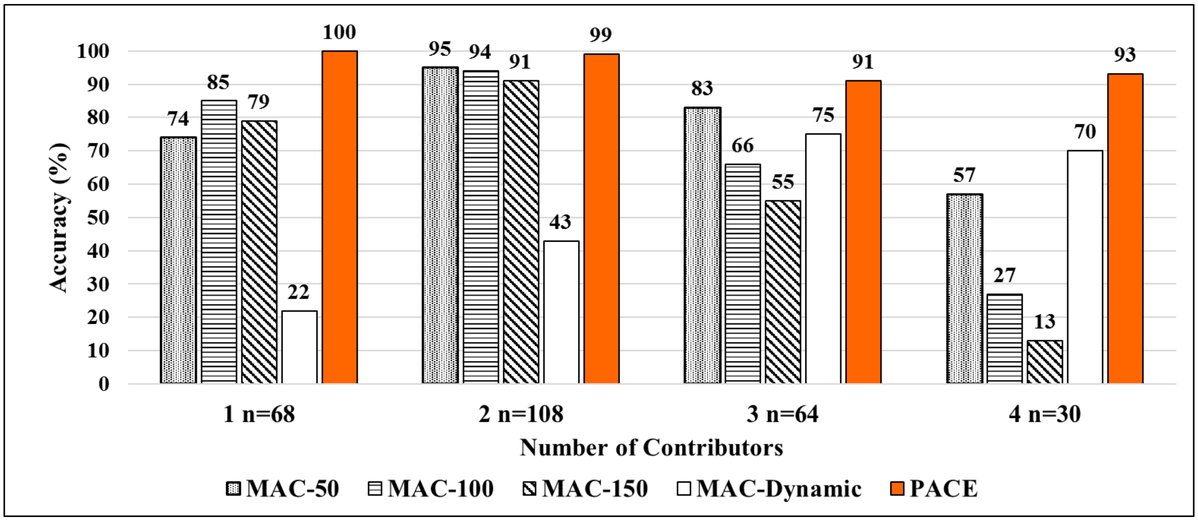

Figure 4: Comparison of PACE-PowerPlex Fusion to MAC. [10]

References

[1] Benschop, C. C., Haned, H., de Blaeij, T. J., Meulenbroek, A. J., & Sijen, T. (2012). Assessment of mock cases involving complex low template DNA mixtures: a descriptive study. *Forensic Science International: Genetics*, 6(6), 697-707.

[2] Bright, J. A., Curran, J. M., & Buckleton, J. S. (2014). The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. *Forensic Science International: Genetics*, 12, 208-214.

[3] Clayton, T. M., Whitaker, J. P., Sparkes, R., & Gill, P. (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1), 55-70.

[4] Butler, J. M. (2014). *Advanced topics in forensic DNA typing: interpretation*. Academic Press.

[5] Paoletti, D. R., Doom, T. E., Krane, C. M., Raymer, M. L., & Krane, D. E. (2005). Empirical analysis of the STR profiles resulting from conceptual mixtures. *Journal of Forensic Science*, *50*(6), JFS2004475-6.

[6] Perez, J., Mitchell, A. A., Ducasse, N., Tamariz, J., & Caragine, T. (2011). Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts. *Croatian medical journal*, *52*(3), 314-326.

[7] Coble, M. D., Bright, J. A., Buckleton, J. S., & Curran, J. M. (2015). Uncertainty in the number of contributors in the proposed new CODIS set. *Forensic Science International: Genetics*, *19*, 207-211.

[8] Egeland, T., Dalen, I., & Mostad, P. F. (2003). Estimating the number of contributors to a DNA profile. *International journal of legal medicine*, *117*(5), 271-275.

[9] Haned, H., Pene, L., Lobry, J. R., Dufour, A. B., & Pontier, D. (2011). Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count?. *Journal of forensic sciences*, *56*(1), 23-28.

[10] Marciano, M. A., & Adelman, J. D. (2017). PACE: Probabilistic Assessment for Contributor Estimation—A machine learning-based assessment of the number of contributors in DNA mixtures. *Forensic Science International: Genetics*, *27*, 82-91.

[11] Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

[12] Duygulu, P., Barnard, K., de Freitas, J. F., & Forsyth, D. A. (2002, May). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision* (pp. 97-112). Springer, Berlin, Heidelberg.

[13] Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

[14] Cho, S. B., & Won, H. H. (2003, January). Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*-Volume 19 (pp. 189-198). Australian Computer Society, Inc.

[15] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 61-74.

[16] Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999, January). When is "nearest neighbor" meaningful?. In *International conference on database theory* (pp. 217-235). Springer, Berlin, Heidelberg.

[17] Bellman, R. (2013). *Dynamic programming*. Courier Corporation.

[18] Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms* (No. VG-1196-G-8). CORNELL AERONAUTICAL LAB INC BUFFALO NY.

[19] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

[20] Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research, 5(Sep), 1089-1105.

[21] Lemaıtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1-5.