

Wing Pattern-Based Classification of the *Rhagoletis pomonella* Species Complex Using Genetic Neural Networks

Chengpeng Bi^{1,2,†‡} Michael C. Saunders^{1,2} Bruce A. McPheron²
cbi@cmh.edu mcs5@psu.edu bam10@psu.edu

¹ Inter-college Operations Research Program, Pennsylvania State University, University Park, PA 16802, U.S.A.

² Department of Entomology, Pennsylvania State University, 501 ASI Building, University Park, PA 16802, U.S.A.

Abstract

The *Rhagoletis pomonella* species complex consists of at least four sibling species. They are highly host specific as larvae, and display great fidelity as adults. The only certain way to identify them is to know the host materials from which they came, because these fruit flies are very similar or identical, and have been especially recalcitrant to morphological separation. In this paper we hypothesize that there is hidden biological information in the wing vein structure in the *pomonella* species group that can be used to distinguish them. Classification of the species complex is modeled via Bayesian and probability neural networks using information on wing size, shape and vein structure. The classification models were optimized through a genetic algorithm by selecting the optimal features and performed well in classifying new specimens. The results have implications for agricultural production and quarantine issues and could be helpful in devising a classification system for rapid identification of certain invasive species at ports of entry.

Keywords: *Rhagoletis pomonella* species complex, Bayes theory, Probability Neural Network, Genetic Algorithm, Pattern Classification

1. Introduction

The *Rhagoletis pomonella* species complex was established by Bush in his 1966 monograph on the *Rhagoletis* fruit flies [1]. This species group, in the taxonomic family, Tephritidae, consists of four sympatric sibling species and three undescribed species [2,3], each of which is restricted to a single host or a group of closely related hosts. One of these species, *R. pomonella*, is comprised of two host races, one on native hawthorns and the other on apples. This host race complex has been the subject of intensive study on modes of speciation [4]. The economic impact of the *Rhagoletis* fruit flies is high in the United States and Canada, primarily because of the widespread infestation of apples by *R. pomonella*. Unfortunately, this species group has been especially recalcitrant to morphological discrimination [2]. Developing a quick and precise method to discriminate among these insect pests is very important for agricultural production and quarantine.

Taxonomic characters found in the tephritid wing are of primary importance in the identification of most genera and species [1,5]. These reside not only in the pattern but also in the arrangement of the veins and shapes of the cells. A right wing image and landmarks of *R. pomonella* are illustrated in Figure 1. Previous quantitative investigations of wing structure in the *R. pomonella* group have been neglected [1]. *Drosophila* wing size and shape were used to study population genetics and answer questions regarding natural selection and conservation genetics although measurements simply consist of wing area or perimeter and lengths of two cross-veins [6].

[†] Current address: Laboratory of Bioinformatics and Intelligent Computing, Children's Mercy Hospitals

[‡] Corresponding author

Bayesian decision theory is a fundamental technique for pattern recognition and classification [7-10]. This approach assumes that patterns possess random characteristics and that they are generated in a random way by some natural phenomena or set of processes. The Bayesian method is based on quantifying the tradeoffs between various classification decisions using probability theory and the costs that accompany such decisions. It assumes that the decision problem is posed in probabilistic terms and that all of the relevant probability values are known. The Bayes decision theory provides a framework for handling the required probability descriptors of the pattern processing problem. It provides statistical methods for classifying patterns into classes based on probabilities of patterns and their features.

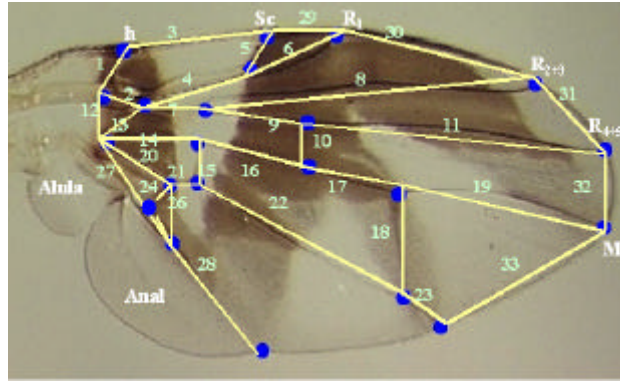


Figure 1: Wing landmarks and segment labels. The solid circles are anatomic landmarks. All the numbers are measured segments and defined as v_i . The set of segments represents the vein structure or venation.

Genetic Algorithms (GAs) [11] are adaptive and robust computational procedures modeled on the mechanics of natural genetic systems. GAs act as a biological metaphor and try to emulate some of the processes observed in natural evolution. They are viewed as randomized, yet structured, search and optimization techniques. GAs are very useful tools in the optimal searching of a large and complex dimension space. There are numerous GA applications in optimization problems such as image processing, medicine, and pattern recognition [12,13].

We hypothesize that there is hidden biological information in the wing vein structure and banding patterns in the *pomonella* species group that can be used to distinguish species. This paper will focus on mining the hidden information in wing structure and banding patterns. The wing variables include quantitative features of vein structure and banding pattern. The objectives of this paper are: (i) to describe basic statistical analysis of the wing variables and wing morphometrics; (ii) to explore the hidden information in the venation and build vein classification models (Linear Fisher Discriminant functions and Probability Neural Network models); and (iii) use a Genetic Algorithm to optimize these models by feature selection (i.e. dimension reduction).

2. Materials And Methods

2.1 Measurements of Wing

Four fruit fly species, *R. pomonella* (Walsh), *R. mendax* Curran, *R. zephyria* Snow, and *R. cornivora* Bush, were used for wing vein measurement. In *R. pomonella*, hawthorn and apple host races were used. All adult fruit fly specimens came from Dr. Bruce McPheron's laboratory collection, identified by rearing from known host materials. The right wing was removed from each specimen and mounted on a microscope slide with 70% ethanol. The wing image was captured using the *PAX-it* system (MIS, Inc, Franklin Park, IL, USA).

Twenty-two landmarks were established to capture wing venation (Figure 1). They are anatomic landmarks consisting of basal points of veins, forks of vein branches or cross-veins, and the endpoints of veins (Figure 1). Each segment line is drawn between adjacent landmarks. A total of thirty-four lines were drawn and labeled on each fly wing (Figure 1). All the lengths of the segment lines were measured using the

PAX-it image processing system, and the vein morphometric data were saved in MS Excel format files. The dataset was divided into male (146 samples) and female (98 samples) wing subsets to permit examination of sexual dimorphism.

For each right wing we have m observations and here m is equal to 34 measurements or feature variables. Let v_i represents the i^{th} observed feature (the vein segment length in mm) with the feature number labeled as in Figure 1. Suppose we have k instances and each instance is an m -dimensional feature vector. The sample matrix size is k by m . Let v_{ij} be any element and thus be the instance i with feature j . Each v_{ij} is subject to the following transformation:

$$x_{ij} = \frac{v_{ij} - \text{mean}(v_j)}{\text{std}(v_j)}, j = 1, \dots, m \quad (1)$$

This is also called z-score standardization. Equation (1) applies to all taxa.

2.2 Bayesian Classification Models

Each individual fly wing is an object of interest. Each object has a set of real-valued features (x) to describe. Here $x = (x_1, x_2, \dots, x_m)^T$ and $x \in \mathbb{R}^m$ ($m = 34$). Objects are classified into n taxa ($n = 5$). Let S denote the set of fruit fly taxa, $S = \{s_1, s_2, \dots, s_n\}$. Given a feature vector or vein pattern x and prior probabilities $P(s_j)$ for all $j \in \{1, 2, \dots, n\}$, we can map vector x to the set S with various posterior probabilities $P(s_j | x)$. Based on Bayesian theorem, we have,

$$P(s_j | x) = \frac{f(x | s_j)P(s_j)}{f(x)}, j = 1, 2, \dots, n, \quad (2)$$

where $f(x|s_j)$ is the conditional joint probability density function for taxon j . The term $f(x)$ represents the unconditional joint probability density function for a vein pattern x . We know that

$$f(x) = \sum_{j=1}^n f(x | s_j)P(s_j), j = 1, 2, \dots, n \quad (3)$$

2.3 Parametric Estimation

Assume that the conditional joint probability density function $f(x|s_j)$ is a multivariate normal density function,

$$f(x | s_j) = \frac{1}{(2\mathbf{p})^{m/2} |\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(x - \mathbf{m}_j)^T \Sigma_j^{-1} (x - \mathbf{m}_j)\right] \quad (4)$$

Suppose $\mathbf{S}_j = \mathbf{S}$ for all $j = \{1, 2, \dots, n\}$, the estimates of \mathbf{S} and \mathbf{m}_j are given by the pooled covariance and host-related mean vector [14]. If we take the logarithm of the above equation (4), the simplified function is called the Fisher linear discriminant function (FLDF) model which gives rise to,

$$d_j(x) = \mathbf{m}_j^T \Sigma^{-1} x - 0.5 \mathbf{m}_j^T \Sigma^{-1} \mathbf{m}_j + \ln[P(s_j)] \quad (5)$$

2.4 Nonparametric Estimation

Based on the Parzen-window approach [8], we give the estimate of probability density function for a given pattern x based on prototypic patterns as,

$$\hat{f}(x | s_j) = \frac{1}{(2\mathbf{p})^{m/2} h_j^m N_j} \sum_{i=1}^{N_j} \exp\left(-\frac{(x - x^{j,i})^T (x - x^{j,i})}{2h_j^2}\right) \quad (6)$$

Here h_j is a smoothing parameter for taxon j and N_j is the total training samples in taxon j . A proper selection of h_j is crucial for obtaining a good estimate of true density. The window-based probability density estimation and optimal Bayesian classification rule can be implemented via a corresponding probability neural network (PNN) [15]. The pseudo-code of the PNN algorithm is as follows,

Algorithm 1: PNN

```

<1> input layer
Given an unknown pattern or feature vector  $\mathbf{x}$ 
<2> pattern layer:  $\mathbf{x}^i$  is the  $i^{\text{th}}$  reference pattern vector
for  $i = 1:N$ 
   $y^i = \mathbf{x}^i \mathbf{x}^T - 0.5(\mathbf{x} \mathbf{x}^T + \mathbf{x}^i (\mathbf{x}^i)^T)$ ;
   $y^i = \exp(y^i/h^2)$ ; % go through activation function
end
<3> summation layer
for  $j = 1:n$ 
   $\text{sum}(j) = 0$ ;
  for all  $i$  in  $\{1, \dots, N\}$  % all instances in the same taxon
     $\text{sum}(j) += y^{(j,i)}$ ;
  end

   $\text{sum}(j) = \frac{\text{sum}(j)}{(2p)^{m/2} h^m N_j}$  % go through activation function
end
pattern  $\mathbf{x}$  belongs to taxon  $j$  with some memberships as
 $\text{membership}(j) = \frac{\text{sum}(j)}{\sum_{j=1}^n \text{sum}(j)}$  for all  $j$  in  $[1,n]$ 

<4> output layer
assign pattern  $\mathbf{x}$  to taxon  $j$  ( $s_j$ ) with the highest membership such that
 $s_j^* = \underset{\text{all } j \in \{1, \dots, n\}}{\text{arg max}} \{ \text{membership}(j) \}$ 
Conclusion: assign  $\mathbf{x}$  to taxon  $j$  with  $\text{membership}(j)$ .

```

The above PNN algorithm was implemented in the matlab language.

2.5 Simulation Parameter Adjustment

Let us set h in range of $[0.1, 5]$ with step length of 0.1. For each h in this range, we run the above PNN algorithm with a Jackknife estimation (discussed below) to evaluate the performance. The optimal h corresponds to the highest performance.

Assume that we have a complete training dataset (patterns) and develop a Bayesian classifier as mentioned previously. The test dataset may be corrupted in particular known ways, for example, a fly wing is partially destroyed and we can only measure part of the required lengths. In this case, we replace the missing values with the corresponding feature averages.

Given a new pattern \mathbf{x} , if some of the feature values are out of bounds, we replace the outliers with the values of lower or upper bounds. Let each feature value x_i have a range of $[x_{low}, x_{up}]$, then we have,

$$x_i = \begin{cases} x_i & x_i \in [x_{low}, x_{up}] \\ x_{low} & x_i < x_{low} \\ x_{up} & x_i > x_{up} \\ \bar{x}_i & \text{otherwise} \end{cases} \quad (7)$$

2.6 Evaluation

For classification problems, it is natural to measure the system performance in terms of the error rate. The classification system predicts the taxon of each pattern of feature vector \mathbf{x} : If it is correctly assigned, it is counted as a success; if not, it is counted as an error. We use the Jackknife or “leave-one-out” estimation method [8,16] to evaluate the performance of the classification models. This is a sample partition and operates so for N times. The jackknife estimate of classifier accuracy is the mean of all those N leave-one-out accuracies. We calculate the observed success rate f based on N leave-one-out instances. The following formula [16] is used to calculate the confidence limits of p :

$$p = \frac{f + \frac{z^2}{2N} \pm z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \quad (8)$$

N is the sample size of the test data and z is the value corresponding to the confidence range. For example if the confidence range is 90%, the z value is at 1.65. If two classifiers fall in the same confidence range of accuracy, their performances are not statistically significant.

For each training set we constructed a confusion matrix \mathbf{C} in which the row index (i) is the true taxon a pattern belongs to and the column index (j) is the taxon predicted by the classifier model. Thus, we have a square matrix $\mathbf{C} = (p_{ij})_{n \times n}$. Each p_{ij} represents the number of instances that taxon i is predicted as taxon j . The number of correctly classified instances is shown on the diagonal. The estimated success rate \hat{p} and its variance are defined by,

$$\hat{p} = \frac{\text{trace}(\mathbf{C})}{\sum_{i=1}^n \sum_{j=1}^n p_{ij}} \quad \text{and} \quad \hat{q} = 1 - \hat{p} \quad (9)$$

$$\text{Var}(\hat{p}) = \frac{\sum_{i=1}^n (\hat{p}_i - \bar{p})^2}{n-1} \quad \text{and} \quad \bar{p} = \frac{\sum_{i=1}^n \hat{p}_i}{n} \quad (10)$$

2.7 Genetic Algorithm for Feature Selection (GAFS)

If we assume that those features that only make a small contribution to an accurate classification are negligible, then those features can be removed. The goal of feature selection is to draw a subset \mathbf{r} from the feature space. When the feature space is large, combinatorial explosion occurs and it is impractical to do an exhaustive search. Genetic algorithms (GA's) are appropriate in handling such cases. A genetic algorithm [17,18,19] is a stochastic process performing a search over a complex and multi-modal space. It is a randomized method in that it utilizes domain knowledge, in the form of objective functions, to perform a directed random search. Here we use GA to perform feature selection and actually perform dimension reduction (DR). The objectives of GAFS are to minimize the generalization error (i.e. improve the performance), keep the training error as low as possible, and reduce the feature dimension. The output of the fitness functions is based on algorithms 1 and 2 (detailed in pseudo-codes) combined with the Jackknife estimation method. The fitness output (testing performance or error rate) directs the optimal search. We use each chromosome to encode a subset of the feature space and thus each locus on a chromosome is simply 0 or 1 representing a feature either present or absent. Thus, each chromosome is an ordered string of binary numbers. Each position on the sequence specifies a feature variable. While decoding we take one string at a time and remove those feature variables with 0's and make a new subset of the feature space. The population size of chromosomes is set to 50 and lasts 100 generations. The mutation rate is 0.001 and the crossover rate is 0.6.

An elitism algorithm used to ensure that the individual with the largest fitness (elite chromosome) in the current population will be passed into the next generation, was implemented. A random number k is

generated within (1, 50) and then the k^{th} individual chromosome is replaced with the current elite chromosome. Fitness is the success rate or performance on the testing dataset by the two functions: FLDF and PNN. The GAFS algorithm was implemented in the matlab language and run with Windows NT/2000 and Unix.

The GAFS model was applied again to the FLDF model with gender (male or female) as our response variable instead of taxon. All the parameter settings for GAFS were the same as above.

The GAFS algorithm is implemented in matlab language and its pseudo-code is outlined as follows,

Algorithm 2: GAFS

```

Set population_size = 50 and generation = 100
Set G = 0 and also set h = 1.2 or 1.6 for PNN model option
Encoding a population of chromosomes
while G < generation
    G = G + 1;
    for i = 1:population_size
        decoding each chromosome and form a new dataset
        for each sample partition of jackknife method
            calculating the fitness using options:
            1) Run FLDF model
            2) Run PNN model
        end
    end
    begin genetic operation
        find the best chromosome and keep it as the elite
        do selection (wheel roulette, Mitchell, 1997)
        do mutation
        do crossover (two-point crossover, Mitchell, 1997)
        form a new population
        do elitism
    end
end
end

```

3. Experimental Results

All algorithms are implemented using matlab. Table 1 and 2 summarize the basic statistics of the training patterns. The minimum and maximum values as well as averages are used to detect outliers and replace missing values. They are also utilized in training and testing the PNN network and the FLDF models.

The jackknife method was used in the PNN model to estimate the success rate on the testing dataset of the male and female groups. We take the optimal h^* corresponding to the highest performance (success rate in Figure 2). The male value of $h^* = 1.2$ is different from that of females, $h^* = 1.6$. In the male group large h values lead to poorer performance.

Thirty-four vein segment lengths (target variables shown in Table 1 and 2) were used to fit to the FLDF and PNN models. The Jackknife method was used to estimate success rates on training and test datasets for male and female groups, and the results are shown in Table 3. In the male group, the FLDF classifier has a 76.7% performance on the testing dataset and its confidence interval is from 70.5% to 82.0%. The PNN classifier has a 63.0% performance on the same test dataset as the FLDF classifier's and its confidence interval ranges from 54.0% to 70.6%. In the female group the PNN classifier has a higher performance at 72.5% (with confidence interval of 66.0%, 78.1%) than the FLDF classifier which performed at 68.4% (with confident interval of 60.2%, 75.6%). The ϵ -substitution success rates of two classifiers on the training dataset for male and female groups are over 95% (Table 3).

Table 4 shows the results of the GAFS models. After 100 generations the feature dimension was reduced from 34-D to 19-D for FLDF models of male and female groups and also decreased from 34-D to 20-D for the male PNN model and to 21-D for the female PNN model. The dimension reduction leads to performance improvement for PNN and FLDF models: The male FLDF and PNN models improved to 84.9% and 70.6%,

respectively, and female models to 83.7% and 78.6%, respectively. The progress of performance improvement by generation for FLDF and PNN models is shown in Figures 3 and 4. We see that after feature selection or dimension reduction is performed, the performance of male and female FLDF models reaches the same level without significant difference (Figure 3a). On the other hand, PNN models still display gender differences, with the female PNN model demonstrating a higher performance than that of males (Figure 3b).

For the FLDF models, we should measure the training error rate while the performance is improving. Figure 3a and 3b show that at the very beginning when dimension is decreasing and performance improving, the training error increases and then as the performance improves further, the training error decreases. The curves also indicated that, given the parameter settings, convergence is reached in 100 generations.

The PNN and FLDF models were optimized by GAFS with the Jackknife estimation. The dimension reduction leading to optimal models was due to noise removal and clearance of some conflicting variables. Another benefit we gain is that dimension reduction dramatically decreases the computing time. The parameters for FLDF male and female models are found in Tables 5 and 6. For the PNN models, the feature dimension of each reference pattern and input pattern is reduced.

Figure 5 shows the progress of searching for the best performance of a gender separating classifier. The peak of success rate was found at 90.89% on testing data with 90% confidence intervals of 87.38%, 93.49%. The re-substitution success rate is 91.4% with standard error of 0.16%. Table 7 shows the estimated parameters of the GSF classifier.

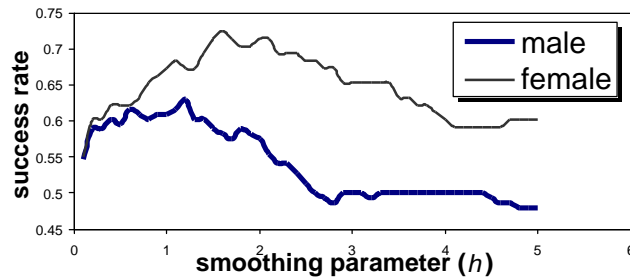
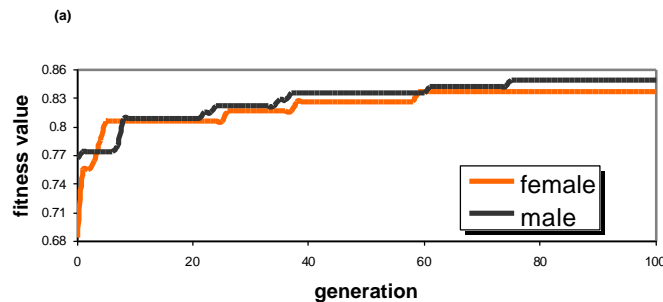


Figure 2: Simulation results on smoothing parameter. The optimal smoothing parameters in PNN model were estimated with Jackknife method. The optimal $h^* = 1.6$ for the females and $h^* = 1.2$ for the males as seen the peaks of the curves.



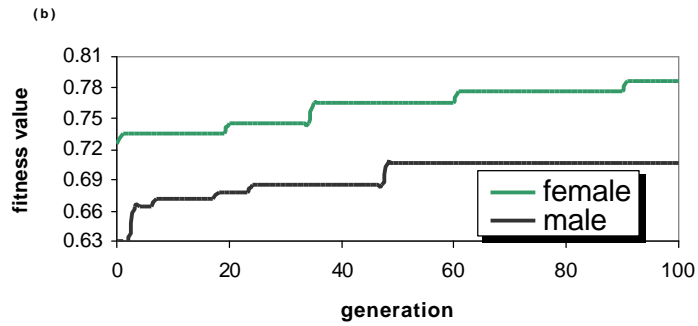


Figure 3: Performance improvement by GAFS generation. (a) GAFS model applied on FLDF models and finally female and male models reach the same fitness value. (b) GAFS applied on PNN models and results show that male and female are different fitness value. The fitness value represents the success rate on test dataset.

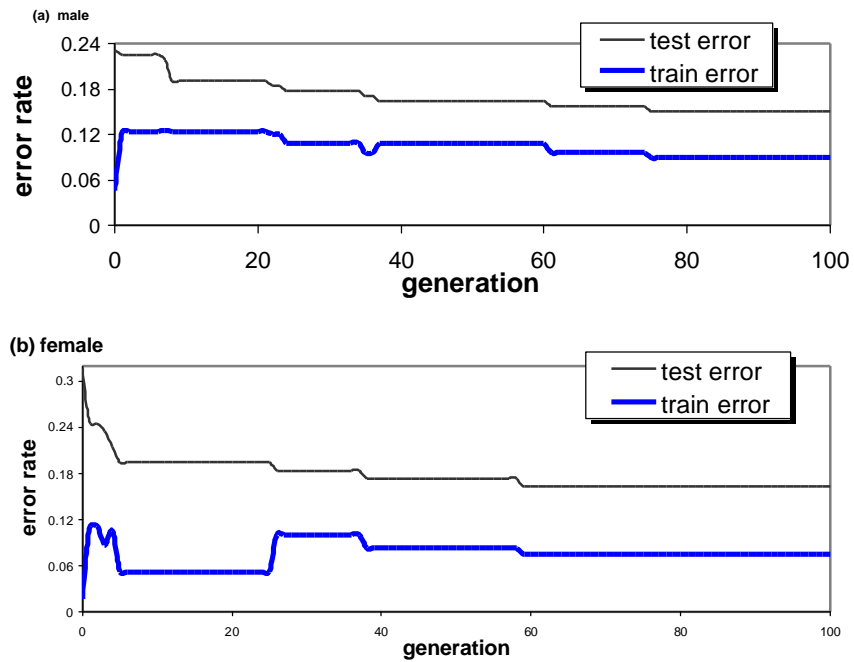


Figure 4: FLDF train error vs test error by GAFS generation. (a) GAFS applied on male FLDF model. (b) GAFS applied on female FLDF model. In both case the error rates on test dataset converge.

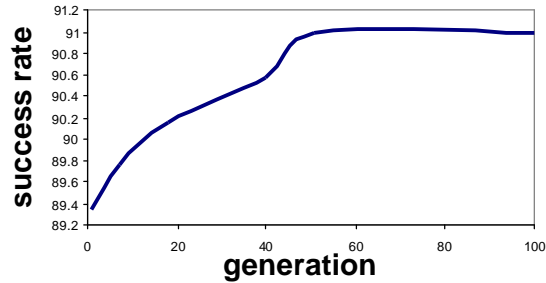


Figure 5: GAFS for gender discriminant function: performance improvement through generations. Jackknife estimation method was used and run 100 generations for GAFS. At generation 51, the success rate on testing data reached the peak with dimension reduction to fourteen feature variables.

4. Discussion

We have shown that there is hidden biological information in the wing vein structure in the *pomonella* species group that can be successfully used to distinguish species. The FLDF models are the best classifiers in terms of performance. The PNN models have the advantage in that these classifiers can memorize the patterns that already have been learned. The results have implications for agricultural production and quarantine issues and could be helpful in devising a classification system for rapid identification of invasive species at ports of entry.

It was found that most of the misclassified instances are among individuals from taxon 2 and 3, likely because they belong to the same species (*R. pomonella*), and they are morphologically almost identical. For example in the male PNN model, there are 9.02% apple race individuals misclassified as hawthorn race and 90.08% are correctly classified. On the other hand 7.1% of the hawthorn race individuals are misclassified as belonging to the apple race and another 3.33% are misclassified as *R. mendax*. The other three taxa have much higher degrees of separation.

In Bayesian modeling we used two methods (parametric and nonparametric) to estimate the conditional probability density functions. In the parametric methods we assume that the forms of probability densities are multivariate Gaussian density functions. Although in practical classifications like our complicated case this assumption is suspect and multi-modal functions are common, the performance shows that this simplified assumption works very well as it has in most cases [7,8]. On the other hand, the Parzen window method approximates the densities by learning from sample data [7,8]. The Parzen method assumes no density functions and its calculation does not need matrix inversion. In sibling species recognition, a covariance matrix usually contains nearly identical columns and thus the matrix is near singular [20]. This is a serious problem when applying the inverse of the matrix in parametric method such as quadratic discriminant functions. The Parzen window approach avoids these problems. But the performance of Parzen window method depends on the quality of prototype or reference patterns. The more reference patterns the better. The computing cost is very high for some large reference patterns. One of PNN model's advantages is the capability of memorizing the training samples. The weak point for neural networks is in over-fitting, resulting in sub-optimal performance on testing samples.

FLDF is a very simple model but very powerful and efficient in terms of performance and computing cost. After applying the GAFS model, the FLDF male and female models reach the same level of performance. The GAFS method was used to perform feature selection and thus reduce the dimension size and improve performance as well. GA's employ a random, yet directed, search for locating the globally optimal solution. They are superior to gradient descent techniques [21] as the search is not biased towards the locally optimal solution. GA's are actually heuristic searching methods and appear to work well in this situation although it is not guaranteed that it will be the best approach. Additionally the globally optimal or sub-optimal solution is not unique because the vein segment variables are correlated to some extent and redundant information also exists. The principal component analysis can make all the variables independent

but the transformed variables (principal components) are very hard to interpret and thus biologically meaningless. The GAFS performed well in feature selection and in extracting the best feature subsets.

Acknowledgments

We thank Drs. Stuart Berlocher, Hong Chen and Dietmar Schwarz for providing the fruit fly specimens. Thanks also go to Dr. Paul Heinemann and Dr. William Roush for their invaluable comments on this paper.

References

- [1] Bush, G. L. 1966. The taxonomy, cytology, and evolution of the genus *Rhagoletis* in North America (Diptera, Tephritidae). *Bulletin of the Museum of Comparative Zoology* 134 (11): 431-562
- [2] Berlocher, S. H. 2000. Radiation and divergence in the *Rhagoletis pomonella* species group: inferences from allozymes. *Evolution* 54: 543-557.
- [3] Feder, JL, Berlocher, SH, Roethele, JB, Dambroski, H, Smith, JJ, Perry, WL, Gavrilovic, V, Filchak, KE, Rull, J, and Aluja, M 2003. Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. *PNAS USA*. 100:10314-10319.
- [4] Berlocher SH, Feder JL. 2002. Sympatric speciation in phytophagous insects: Moving beyond controversy? *Annual Review of Entomology*. 47:773–815.
- [5] Foote, R., F. Blanc & A. Norrbom 1993. *Handbook of the Fruit Flies (Diptera: Tephritidae) of America North of Mexico*. Cornell University Press. Ithaca, New York. 571 pp.
- [6] Gilchrist, A. S. and L. Partridge 2001. The constraining genetic architecture of wing size and shape in *Drosophila melanogaster*. *Heredity* 86: 144-152.
- [7] Cios, K., W. Pedrycz, and R. Swiniarski 1998. *Data Mining Methods for Knowledge Discovery*. Kluwer Academic Publishers. Norwell, MA. 495 pp.
- [8] Duda, R. O., P. E. Hart and D. G. Stork 2001. *Pattern Classification*. John Wiley & Sons, Inc. New York. 654 pp.
- [9] Fu, K. S. 1982. *Syntactic Pattern Recognition and Applications*. Prentice Hall. Englewood Cliffs, NJ. 596 pp.
- [10] Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press. New York. 403 pp.
- [11] Holland, J. H. 1975. *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Arbor. 183 pp.
- [12] Mitchell, M. 1998. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA. 206 pp.
- [13] Pal, S. K. and P. P. Wang 1996. *Genetic Algorithms for Pattern Recognition*. CRC Press. Boca Raton, FL. 314 pp.
- [14] Johnson, R. and D. Wichern 1998. *Applied multivariate statistical analysis*. Prentice Hall. Upper Saddle River, NJ. 816 pp.
- [15] Specht, D. F. 1990. Probabilistic neural networks. *Neural Networks*, 3(1): 109-118.

- [16] Witten, I. H. & E. Frank 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, San Francisco, CA. 371 pp.
- [17] Back, T. 1996. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming and Genetic Algorithm*. Oxford University Press. Oxford. 314 pp.
- [18] Davis, L. 1991. *Handbook of genetic algorithms*. Van Nostrand Reinhold, New York. 385 pp.
- [19] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA. 412 pp.
- [20] Golub, G. H. & C. F. Van Loan 1989. *Matrix computations*. The Johns Hopkins University Press. Baltimore, MD. 694 pp.
- [21] Tsoukalas, L. H. and R. E. Uhrig 1997. *Fuzzy and Neural Approaches in Engineering*. John Wiley & Sons, Inc. New York, NY. 587 pp.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉
Mean	0.254	0.239	0.862	0.662	0.232	0.601	0.388	1.980	0.614
Std	0.035	0.043	0.095	0.079	0.034	0.073	0.058	0.197	0.087
Min	0.170	0.146	0.618	0.467	0.163	0.453	0.236	1.455	0.361
max	0.367	0.404	1.102	0.826	0.328	0.790	0.534	2.452	0.824
	V ₁₀	V ₁₁	V ₁₂	V ₁₃	V ₁₄	V ₁₅	V ₁₆	V ₁₇	V ₁₈
Mean	0.231	1.728	0.275	0.329	0.587	0.201	0.625	0.623	0.522
Std	0.026	0.173	0.036	0.040	0.070	0.026	0.079	0.067	0.056
Min	0.152	1.300	0.176	0.215	0.419	0.137	0.450	0.465	0.381
max	0.289	2.083	0.373	0.434	0.754	0.254	0.790	0.804	0.646
	V ₁₉	V ₂₀	V ₂₁	V ₂₂	V ₂₃	V ₂₄	V ₂₅	V ₂₆	V ₂₇
Mean	1.123	0.475	0.143	1.348	0.218	0.160	0.204	0.275	0.646
Std	0.123	0.056	0.028	0.147	0.028	0.026	0.035	0.038	0.083
Min	0.816	0.311	0.078	0.992	0.151	0.081	0.129	0.152	0.434
max	1.405	0.623	0.216	1.652	0.293	0.226	0.283	0.365	0.864
	V ₂₈	V ₂₉	V ₃₀	V ₃₁	V ₃₂	V ₃₃	V ₃₄	area	Perim.
Mean	0.715	0.457	1.190	0.531	0.420	1.078	1.078	3.614	7.565
Std	0.083	0.058	0.112	0.065	0.053	0.132	0.115	0.684	0.752
Min	0.506	0.328	0.926	0.371	0.293	0.755	0.779	1.969	5.624
max	0.886	0.586	1.465	0.699	0.558	1.317	1.307	5.106	9.055

Table 1 Summary of basic statistics of male fly vein measurements*.

*Note that the total sample size $N = 146$. This is overall summary of the species group.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉
mean	0.316	0.274	1.031	0.796	0.280	0.730	0.476	2.336	0.756
std	0.040	0.040	0.114	0.088	0.039	0.087	0.072	0.224	0.105
min	0.225	0.168	0.761	0.585	0.178	0.544	0.323	1.790	0.470
max	0.402	0.359	1.247	0.991	0.347	0.916	0.616	2.807	0.952
	V ₁₀	V ₁₁	V ₁₂	V ₁₃	V ₁₄	V ₁₅	V ₁₆	V ₁₇	V ₁₈
mean	0.277	2.051	0.330	0.385	0.707	0.244	0.761	0.714	0.635
std	0.031	0.199	0.045	0.048	0.088	0.028	0.100	0.075	0.066
min	0.191	1.601	0.231	0.276	0.503	0.168	0.533	0.500	0.458
max	0.342	2.496	0.440	0.513	0.878	0.305	0.959	0.910	0.763
	V ₁₉	V ₂₀	V ₂₁	V ₂₂	V ₂₃	V ₂₄	V ₂₅	V ₂₆	V ₂₇
mean	1.344	0.571	0.176	1.619	0.255	0.193	0.258	0.338	0.788
std	0.150	0.067	0.035	0.180	0.034	0.026	0.039	0.043	0.096
min	0.988	0.397	0.082	1.190	0.190	0.133	0.170	0.223	0.560
max	1.616	0.722	0.251	1.954	0.371	0.262	0.348	0.437	0.995
	V ₂₈	V ₂₉	V ₃₀	V ₃₁	V ₃₂	V ₃₃	V ₃₄	area	perimt
mean	0.858	0.554	1.391	0.666	0.501	1.262	1.303	5.202	9.053
std	0.109	0.067	0.125	0.085	0.052	0.148	0.144	0.959	0.894
min	0.602	0.420	1.122	0.469	0.346	0.898	0.948	3.031	6.902
max	1.102	0.701	1.673	0.842	0.625	1.560	1.577	7.227	10.731

Table 2 Summary of basic statistics of female fly vein measurements*.

*Note that total sample size $N = 98$. This is overall summary of the species group.

Success rate	Male venation		Female venation	
Variables	34		34	
Dataset	Training	Testing	Training	Testing
FLDF model	0.954± 0.006†	0.767[0.705,0.82 0]‡	0.983±0.00 5†	0.684[0.602, 0.756]‡
PNN model**	1.00± 0	0.630[0.54,0.706]	1.00± 0	0.725[0.66,0.781]

Table 3 Jackknife estimation of performance for FLDF and PNN models*.

*Note that male group with 146 samples and female with 98 samples; for PNN models, the training error rates are always 0.0 due to the z-score transformation and memory of training samples. ** for PNN models, smoothing parameter $h = 1.6$ for female group and $h = 1.2$ for male group; † ± std; ‡ 90% confidence interval.

Success rate	Male venation		Female venation	
Variables	19/20		19/21	
Dataset	Training	Testing	Training	Testing
FLDF model	0.911± 0.0062†	0.849[0.794, 0.892]‡	0.925± 0.009†	0.837[0.767, 0.889]‡
PNN model**	1.00± 0	0.706[0.64, 0.764]	1.00± 0	0.786[0.71, 0.846]

Table 4 Jackknife estimation of optimal performance for GAFS models*

*Note that male group with 146 samples and female with 98 samples; for PNN models, the training error rates are always 0.0 due to the z-score transformation and memory of training samples. ** for PNN models, smoothing parameter $h = 1.6$ for female group and $h = 1.2$ for male group; † ± std; ‡ 90% confidence interval.

Taxon**	1	2	3	4	5
Constant	-140.14	-195.15	-195.58	-146.54	-130.07
v1	-186.19	-184.67	-169.79	-173.00	-169.19
v2	-3.98	6.45	-4.27	-1.97	3.46
v6	6.75	14.07	-21.13	-13.84	-6.04
v8	47.22	21.61	24.31	53.02	63.38
v9	-44.44	-10.54	-35.42	-46.67	-57.64
v11	72.98	81.49	93.59	50.26	15.67
v14	2.22	4.73	8.74	20.57	-5.96
v18	8.35	-60.43	-46.50	-6.46	37.61
v19	29.90	55.04	73.61	22.12	42.47
v21	145.77	158.45	150.81	76.42	142.43
v23	20.50	-44.62	-43.61	44.07	40.23
v25	1.48	11.47	-30.08	-3.85	-26.27
v26	-125.06	-160.01	-115.83	-174.66	-107.81
v28	-45.21	-10.85	-34.09	-39.04	-20.25
v29	20.45	56.02	55.12	42.28	-4.92
v31	27.10	12.62	20.63	61.15	51.23
v32	-35.80	-60.82	-57.65	17.14	-55.32
v33	74.69	99.33	84.03	58.63	40.51
v34	59.69	95.53	99.97	88.76	115.63

Table 5 The coefficients of FLDF male classifier model (Equation 5)*

*Note that male group has 146 samples and calculation is based on original dataset. **taxon 1, 2, 3, 4 and 5 represents *R. mendax*, *R. pomonella* apple and hawthorn races, *R. zephyria* and *R. cornivora*.

Taxon**	1	2	3	4	5
Constant	-229.12	-319.03	-323.29	-226.60	-212.34
v2	275.63	339.55	349.61	229.00	203.73
v3	-279.18	-340.85	-370.95	-298.94	-244.24
v4	241.29	277.97	301.52	238.36	184.05
v6	207.04	277.50	259.62	226.34	138.56
v8	-74.93	-135.76	-104.20	-83.57	-55.58
v9	-11.79	58.60	19.71	-41.31	-21.19
v10	-63.98	-122.45	-157.60	-78.19	-8.22
v12	-81.84	-4.86	-23.59	-46.08	-171.53
v13	266.78	274.22	314.51	248.77	239.81
v15	143.29	48.75	120.64	132.62	304.23
v19	108.44	154.49	152.59	93.78	99.91
v20	75.92	110.62	91.15	110.28	53.80
v22	91.79	86.99	103.27	110.52	87.91
v23	42.77	20.34	27.43	88.04	67.25
v26	-40.17	-86.69	-89.39	-72.97	-26.16
v28	37.65	68.86	65.69	-4.87	21.40
v30	107.92	146.58	129.78	130.65	105.44
v32	33.43	24.20	12.04	93.30	22.64
v34	-58.96	-39.04	-51.06	-56.14	-27.13

Table 6 The coefficients of FLDF female classifier model (Equation 5)*

*Note that female group has 98 samples and calculation is based on original dataset. **taxon 1, 2, 3, 4 and 5 represents *R. mendax*, *R. pomonella* apple and hawthorn races, *R. zephyria* and *R. cornivora*.

Coefficients	female	male
Constant	-90.58	-63.91
v5	-73.04	-57.49
v10	211.68	179.84
v13	73.37	71.63
v14	-53.27	-44.08
v15	155.26	129.52
v16	5.19	6.18
v17	129.16	119.87
v18	122.59	85.64
v19	85.26	69.13
v21	36.34	18.59
v26	-82.25	-79.99
v28	-97.35	-85.44
v31	49.77	27.39
v33	-81.13	-54.80

Table 7 The coefficients of GSF classifier (Equation 5)*

*Note that total sample size is 244. Calculation is based on original dataset. Five taxa are included: *R. mendax*, *R. pomonella* apple and hawthorn races, *R. zephyria* and *R. cornivora*.