USE OF ASSORTED METHODS FOR PREDICTING SUPERIOR SOY BEAN VARIETIES
• Background
• Three techniques
  • Clustering methods
    • K-means
    • Fuzzy C-means
  • Kalman filters
  • Parenclitic networks
    • Boccaletti’s exampled
    • Centrality Measures: Eigenvector centrality
• Goal: Develop a model to predict seed variety performance in soybeans
  • Specifically: to use available genotypic data to create such a model
• Our data are the yield data of soybean growth experiments
  • Include: soil, weather, and location variables along with yield as well as genotypic data for some varieties
• Genotypic data:
  • 2163 genetic markers characterized by single nucleotide polymorphisms (SNPs)
    • SNP: a single base-pair difference in the DNA sequence of unique varieties
    • Given in the form of categorical data where letters represent the different nucleotides
      • (A) adenine, (C) cytosine, (G) guanine, (T) tyrosine, (N) data unavailable
    • Soybeans are diploid organisms meaning each SNP is classified by two nucleotides
  • EX: Marker ID: $X_1 X_5 X_6 X_8$
    • Classification: G|A C|C G|A

CLUSTERING METHODS

Static method
K-MEANS

• Goal: find clusters in the data, with the numbers of groups represented by the variable K

• Algorithm
  • Begins with initial estimate for the K centroids
  • Iterative refinement
    • Each data point is assigned to its nearest centroid using dissimilarities as a comparison measure
    • The centroids are recomputed using the mean (or mode in the case of categorical data) of all data points assigned to that centroids cluster
  • Repeat
K-MEANS

- Applied to our data
  - Clustering the Genotypic information
    - The information given is categorical for each genetic marker
      - Suggests a necessity for using K-modes instead of K-means
      - K-modes are a variation of K-means where mode is used as the distance measure
      - Knowing which cluster a newly sequenced variety belongs to could be used to predict its success based on the average success of the members of the given cluster
  - Other applications
    - Use K-means to cluster based on soil and weather variables
FUZZY C-MEANS

• A version of K-means where each data point can belong to two or more clusters
  • The algorithm is very similar to that of K-means with the added complication of membership coefficients being assigned to each data point instead of a single cluster classification.
  • Data are bound to each cluster by means of a Membership Function.
  • Some varieties in the genetic data are likely to be closely related to more than one family making C-means a more accurate clustering method.
  • If a new variety can be assigned to more than one cluster, it could potentially give more information for the viability of the new variety.
    • For example: a new variety could be classified as halfway belonging to a cluster previously determined to contain varieties that perform well in dry climates and halfway belonging to a cluster previously determined to contain varieties that perform well in soils that have high pH balance.
    • As new varieties are being created and sequenced, it is more likely that they will belong to more than one cluster as they could be descendants from parental types belong to two separate clusters.
A dynamic method to use incoming new data to improve a model
KALMAN FILTERS

- Begin with a prediction and an error estimate
  - \( \hat{x}_k \): prediction
  - \( \hat{x}_{k-1} \): best estimate from previous iteration
  - \( A \): prediction/transition matrix
  - \( B \): control input matrix
  - \( u_k \): control inputs
  - \( P_k \): error estimate
  - \( P_{k-1} \): error estimate from previous iteration
  - \( Q \): Process noise covariance matrix

- Combine the prediction and a given measurement
  - \( K_k \): Kalman Gain, used to make a linear combination that is unbiased and has minimum error
  - \( H \): transformation matrix
  - \( R \): measurement noise
  - \( z_k \): measurement
• In terms of our data:
  • Create a prediction model based on data from the check varieties in the early years ($\hat{x}_k$)
  • Use data from the next data set, let’s say from a given year, to create a prediction model based solely on data from this year. ($z_k$)
  • Compare the prediction model created from the data set before and the prediction model created from the data set of the single year being considered.
  • Combine the two prediction models to make a new, hopefully more accurate prediction model. ($\hat{x}_k$)
  • Using the latest prediction model, repeat the process with data from the next year.
• Goal: To reduce the error in the prediction model to white noise.
PARENCLITIC NETWORKS

A static method to determine variations in the data
STEFANO BOCCALETTI’S EXAMPLE

• Objective: identify genes orchestrating plants response under stress
• Set up
  • Expression level of 1,922 genes
  • 8 different abiotic stresses
  • 6 moments of time
FORMING THE PARENCLITIC NETWORK

• Nodes:
  • Also known as features
  • 1,922 genes being expressed
• Subjects
  • Each plant at each time step
• Note: each time step is treated as a separate set of data unrelated to the other time steps
FORM A PARENCLITIC NETWORK PER DATA SET

- Start with data set 1
- Create a reference model
  - Using data from all the other data sets, excluding data from time step 1
  - Form pairwise relationships
    - Look at gene i
      - Form a regression model (specific to this case) for expected expression level of gene i given gene j
    - Repeat for every pair of genes
- For time step i, weight the link between every pair of genes based on the difference between the real amount expressed and the expected expression given by the reference model
Create a parenclitic network for all 6 time steps
  - Each network begins as an all-to-all connected network
  - Thresholding is introduced onto the edges, eliminating those with values below a certain level in order to obtain a regular network

Calculation of Centrality for each node
  - This measure is what ultimately determines the importance of each node within the network
  - One method:
    - Eigenvector centrality

Result:
  - The gene with the highest centrality measure stands out as potentially important for each specific abiotic stress
  - They were able to determine which genes were expressed differently under certain conditions and which were most important in resisting specific abiotic stressors
One possible method for determining the centrality of each node in the Parenclitic network

Basic idea: Gives each node a centrality score proportional to the sum of the scores of its neighbors

This method increases a node’s importance in a network if it has connections to other nodes that are themselves important

Method:

- Begins by giving each node an initial centrality score
- Updates each node’s centrality score

\[ x_{1}' = \sum_j A_{ij}x_j \]
- \( x_{1}' \): updated centrality score for node \( I \)
- \( A_{ij} \): Adjacency matrix
- \( x_j \): centrality score for all other nodes
- Process is repeated \( t \) times resulting in a vector of centralities

\[ x(t) = A^t x(0) \]
\[ x(0) = \sum_i c_i v_i \]
- \( v_i \): eigenvectors of the adjacency matrix
- \( c_i \): constants

As \( t \) goes to infinity, the leading eigenvector of the adjacency matrix becomes proportional to the limiting vector of centralities, i.e., for \( k_1 \) being the largest eigenvalue,

\[ x_I = k_1^{-1} \sum_j A_{ij}x_j \]
Problem with Eigenvector centrality: In a directed network, if a node is given a centrality score of 0, any node connected to this node by an ingoing edge will also be given a score of 0 and so on.

Page Rank centrality can be used to adjust for this potential problem.

\[ x_j = \alpha \sum_i A_{ij} \frac{x_i}{k_j^{out}} + \beta \]

- \( \alpha, \beta \) : constants
- \( k_j^{out} \) : out-degree for node j; artificially set equal to one for any nodes with zero outgoing edges

This method solves the eigenvector problem and keeps nodes from being scored too highly if they are connected to particularly highly scored nodes.
OUR MODEL

• Given gene expression
  • If gene expression is available, we could create a similar model using different years as different data sets, varieties as the subjects, and genes as the features
  • Instead of time steps we may be able to separate the networks into location groups and determine which genes are being expressed more differently in various locations (under different conditions)

• Without gene expression
  • One idea: create Parenclitic network of “high yield” varieties, weight them based off of a distance measure created for the differences in their genomes, potentially the probability two classifications for specific gene locations would appear together, use the network to determine if a new variety would be classified as “high yield” based on the weight of its links
  • In this case the nodes, or features, would be the varieties themselves
REFERENCES

• http://stanford.edu/~cpiech/cs221/handouts/kmeans.html
• http://bilgin.esme.org/BitsAndBytes/KalmanFilterforDummies
• Evenson, G (2009), Data Assimilation, The Ensemble Kalman Filter
• Newman, M. E. J. (2012), Networks
• Boccaletti, S. et al., (2014), Parenclitic networks: uncovering new functions in biological data