Decision Trees and Random Forests

Yuzhen Ye

School of Informatics, Computing and Engineering

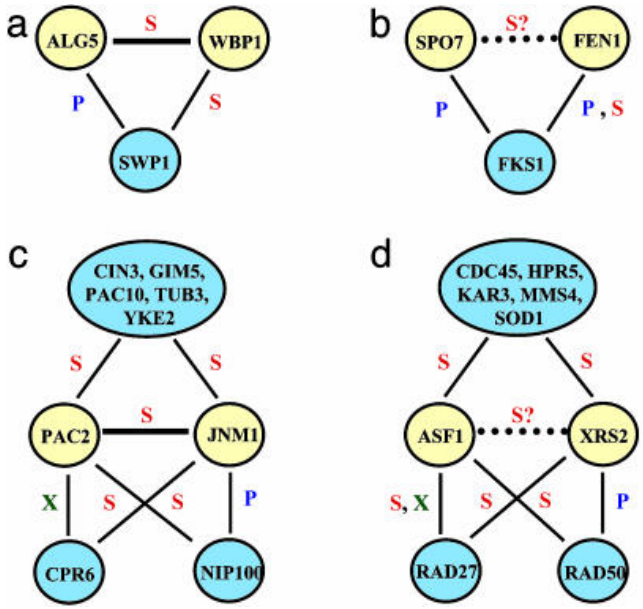
Indiana University, Bloomington

Contents

- Biological problem: SSL prediction
- Decision trees
  - Which attribute is the best?
  - Information gain
- Decision tree variants
  - Random forests
- References

Synthetic sick and lethal (SSL) genetic interactions

- Synthetic sick and lethal (SSL) genetic interactions between genes A and B occur when the organism exhibits poor growth (or death) when both A and B are knocked out but not when either A or B is disabled individually.



Gene-pair relationships. (*a* and *b*) Known (*a*) and predicted (*b*) SSL gene pairs from the highest-scoring leaf of the decision tree. (*c* and *d*) Known (*c*) and predicted (*d*) SSL gene pairs from the third-highest-scoring leaf. P, physical interaction; S, synthetic sick or lethal interaction; X, correlated mRNA expression.

Ref: Proc. Natl. Acad. Sci. USA. 2004;101:15682–15687

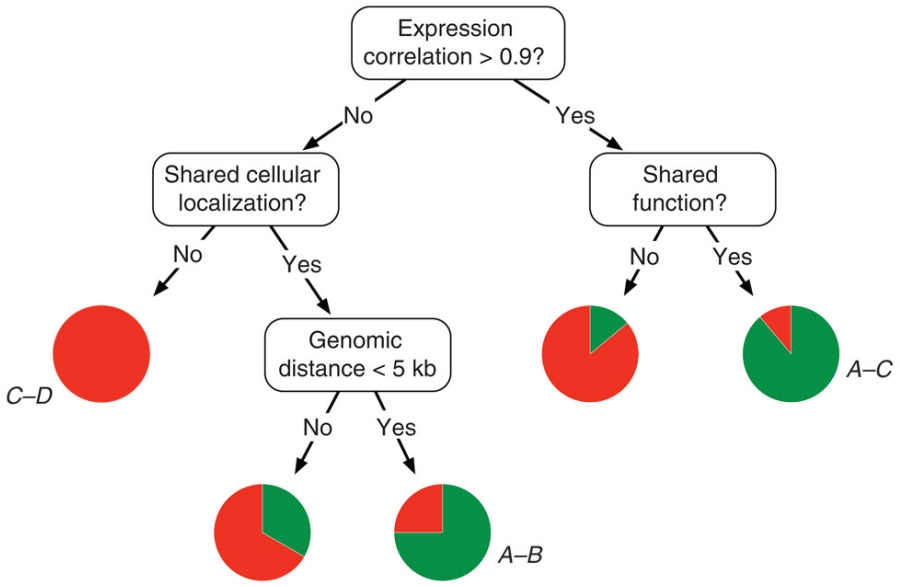
Decision trees

- A decision tree classifies data items by posing a series of questions about the features associated with the items. Each question is contained in a node, and every internal node points to one child node for each possible answer to its question. The questions thereby form a hierarchy, encoded as a tree.

a

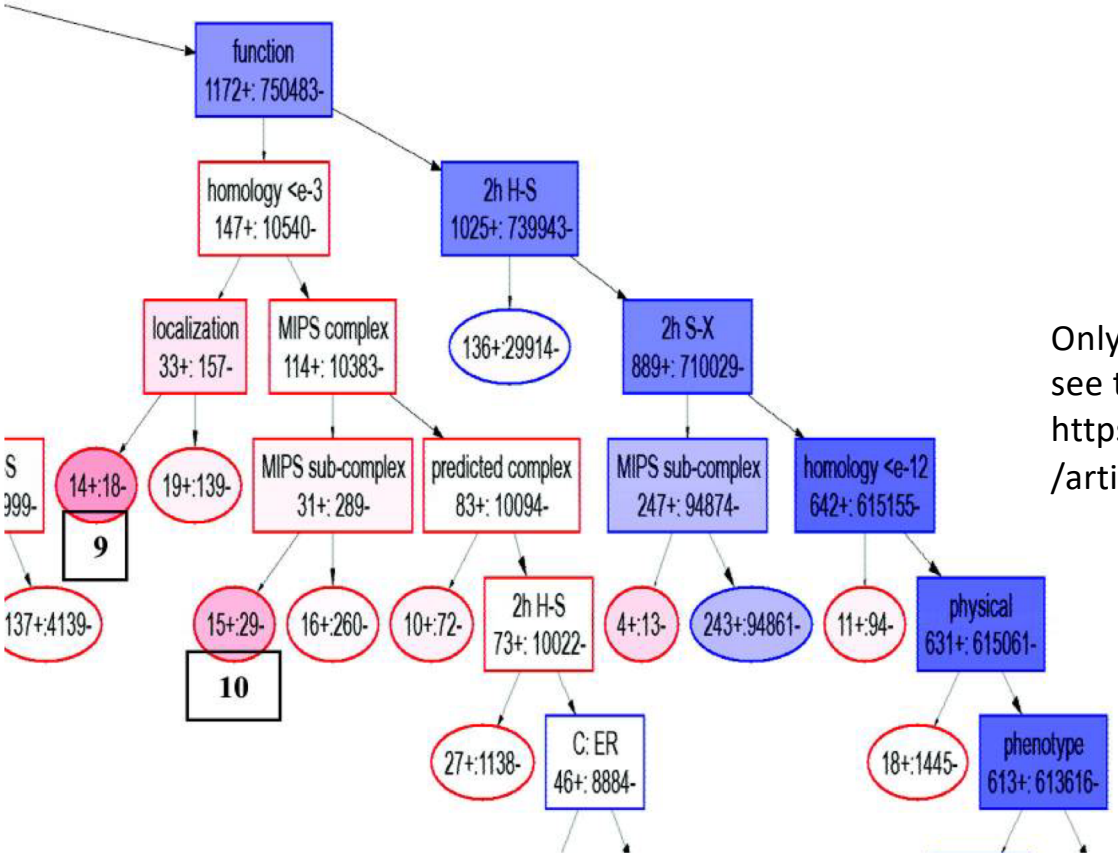
Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb

b



Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701298/

The decision tree for SSL prediction



Only part of the tree is shown here; see the tree at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC524818/

Decision tree reconstruction

- Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down greedy search through the space of possible decision trees.
- At each node, a question is asked “which attribute should be tested here?”
- To answer this question, each attribute is evaluated to determine how well it alone classifies the training examples

Which attribute is the best classifier?

Information gain is often used to measure the "value" of an attribute, which is computed as,

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

Where S is the collection of examples, A is an attribute, $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v. In other words, information gain measures the increase of homogeneity (reduction of impurity) caused by partitioning of the examples according to a given attribute. Here impurity is measured as Entropy, which is computed as

$$Entropy(p) = \sum_i -p_i log_2 p_i \quad (2)$$

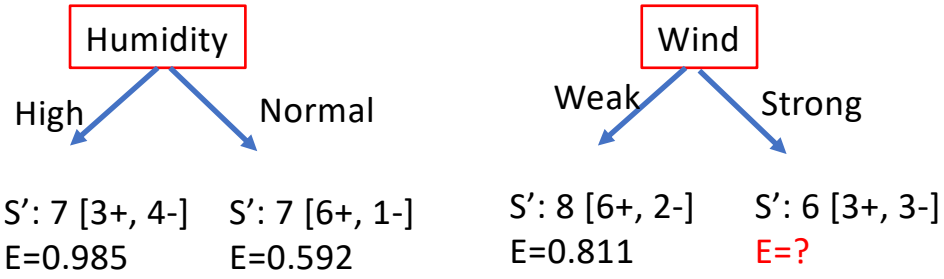
The information gain in layman's language,
Information Gain = Entropy(Parent) – [Average Entropy(Children)]

Ref: Tom Mitchell

See an example

S: 14 [9+, 5-] (9 play, 5 don't play)

$E=-[9/14 \log_2(9/14)+5/14 \log_2(5/14)]=0.94$



Gain(S, Humidity)

=0.940 – [(7/14)*0.985 + (7/14)*0.592]

=0.151

Gain(S, Wind)=?

Is Date is a good attribute?

Date	Outlook	Temperature	Humidity	Wind	Play
5/1	Sunny	Hot	High	Weak	No
5/6	Sunny	Hot	High	Strong	No
5/15	Overcast	Hot	High	Weak	Yes
6/1	Rain	Mild	High	Weak	Yes
6/4	Rain	Cool	Normal	Weak	Yes
6/7	Rain	Cool	Normal	Strong	No
6/11	Overcast	Cool	Normal	Strong	Yes
7/1	Sunny	Mild	High	Weak	No
9/1	Sunny	Cool	Normal	Weak	Yes
9/2	Rain	Mild	Normal	Weak	Yes
9/4	Sunny	Mild	Normal	Strong	Yes
9/7	Overcast	Mild	High	Strong	Yes
9/20	Overcast	Hot	Normal	Weak	Yes
9/21	Rain	Mild	High	Strong	No

Play Tennis Data

Issues in decision tree learning

- Overfitting
  - A hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the unseen instances
  - Happens in other ML approaches
- Approaches to avoid overfitting in decision tree learning.
  - Stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data; e.g., to stop splitting when no question increases the purity of the subsets more than a small amount.
  - Allow the tree to overfit the data, and then post-prune the tree. This can be done by collapsing internal nodes into leaves if doing so reduces the classification error on a held-out set of training examples; or removing nodes in an attempt to explicitly balance the complexity of the tree with its fit to the training data.
  - Post-pruning overfit trees has been found to be more successful in practice
- Training and validation set approach
  - Cross-validation on left-out training examples should be used to ensure that the trees generalize beyond the examples used to construct them.
  - It is important that the validation set be large enough to itself provide a statistically significant sample of the instances.
  - A common heuristic is to withhold one-third of the available examples for the validation set, using the other two-thirds for training.

Alternative measures for selecting attributes

- Gain ratio
  - The Date attribute would separate the training examples (Play Tennis) into very small subsets, so it would have a very high information gain relative to the training examples; but it would be a very poor predictor of the target function over unseen instances
  - The gain ratio measure penalizes attributes such as Date by incorporating a term, called split information, which is sensitive to how broadly and uniformly the attribute splits the data
- Distance-based measure

Random forests

- Many different decision trees are grown by a randomized tree-building algorithm (involving two randomizations).
  - The training set is **sampled with replacement** to produce a modified training set of equal size to the original but with some training items included more than once.
  - When choosing the question at each node, only a small**, random subset of the features** is considered.
  - Each run may result in a slightly different tree.
- The predictions of the resulting ensemble of decision trees are combined by taking the most common prediction.

How many trees in a random forest?

- This work used 29 real medical data to test the impacts of using different numbers of trees
- Results show that sometimes a large number of trees in a forest only increases its computational time, and has no significant performance gain.
- The authors suggest a range between 64 and 128 trees in a forest, to obtain a good balance between AUC, processing time, and memory usage. If the total number of attributes is small, a fewer number of trees (from 8 trees or more) should be used.
- Ref: International Workshop on Machine Learning and Data Mining in Pattern Recognition. MLDM 2012: pp 154-168

Cons and pros of using decision trees

- Pros
  - Decision trees are sometimes more interpretable than other classifiers such as neural networks and support vector machines because they combine simple questions about the data in an understandable way.
  - Decision trees are flexible enough to handle items with a mixture of real-valued and categorical features, as well as items with some missing features.
  - They are expressive enough to model many partitions of the data that are not as easily achieved with classifiers that rely on a single decision boundary (such as logistic regression or support vector machines).
- Cons
  - Small changes in input data can sometimes lead to large changes in the constructed tree.
  - Even data that can be perfectly divided into classes by a hyperplane may require a large decision tree if only simple threshold tests are used.

Knowledge discovery

- From the knowledge discovery perspective, the ability to track and evaluate every step in the decision-making process is one of the most important factors for trusting the decisions gained from data-mining methods.
- Decision trees, along with rule-based classifiers, represent a group of classifiers that perform classification by a sequence of simple, easy-to-understand tests whose semantics are intuitively clear to domain experts
- Although current state-of-the art classifiers (e.g. SVM) or ensembles of classifiers (e.g. Random Forest) significantly outperform classical decision tree classification models in terms of classification accuracy or other classification performance metrics, they are not suitable for knowledge discovery process.
- In most cases the final decision trees will be presented to domain experts for evaluation of extracted knowledge – i.e., rules that can be derived from a decision tree.

Ref: Stiglic G, Kocbek S, Pernek I, Kokol P (2012) Comprehensive Decision Tree Models in Bioinformatics. PLoS ONE 7(3): e33812. doi:10.1371/journal.pone.0033812

Algorithms/tools

- ID3
- C4.5
- C5.0
- CART
- R Package 'randomForest'

References

- **What are decision trees?** Nat Biotechnol. 2008 Sep; 26(9): 1011–1013.
- Machine Learning by Tom Mitchell