Feature relevance estimation by evolving probabilistic dependency networks with weighted kernel machines

Nestor Rodriguez [1]
Advisor: Sergio A. Rojas, PhD.

May 24, 2011

[1] Engineering School, District University of Bogota, Colombia

Introduction

- Feature Subset Selection (FSS) for supervised learning.

Introduction

- Feature Subset Selection (FSS) for supervised learning.
  - Identifies relevant features for building robust learning models.
  - Allows domain experts to get focused on the relevant part of a problem.
  - Increase the generalization capabilities for predictor models in classifications task.

Introduction

- Feature Subset Selection (FSS) for supervised learning.
  - Identifies relevant features for building robust learning models.
  - Allows domain experts to get focused on the relevant part of a problem.
  - Increase the generalization capabilities for predictor models in classifications task.
- Estimation of distribution algorithms

Introduction

- Feature Subset Selection (FSS) for supervised learning.
  - Identifies relevant features for building robust learning models.
  - Allows domain experts to get focused on the relevant part of a problem.
  - Increase the generalization capabilities for predictor models in classifications task.
- Estimation of distribution algorithms
  - Finds optimal solutions for complex search problems.

Introduction

- Feature Subset Selection (FSS) for supervised learning.
  - Identifies relevant features for building robust learning models.
  - Allows domain experts to get focused on the relevant part of a problem.
  - Increase the generalization capabilities for predictor models in classifications task.
- Estimation of distribution algorithms
  - Finds optimal solutions for complex search problems.
  - The solution is represented by probability distribution model.

Introduction

- Feature Subset Selection (FSS) for supervised learning.
  - Identifies relevant features for building robust learning models.
  - Allows domain experts to get focused on the relevant part of a problem.
  - Increase the generalization capabilities for predictor models in classifications task.
- Estimation of distribution algorithms
  - Finds optimal solutions for complex search problems.
  - The solution is represented by probability distribution model.
  - Could be used as a FSS to find relevant variables.

Introduction

- Kernel Classifiers

Introduction

- Kernel Classifiers
  - Kernel classifiers have recently emerged as powerful high-accuracy classifiers with robust generalization abilities.
  - Use *Kernel* functions as a mappings of similarity measures in a transformed space where nonlinearities can be solved linearly.
  - Two widely used kernel functions are:

Introduction

- Kernel Classifiers
  - Kernel classifiers have recently emerged as powerful high-accuracy classifiers with robust generalization abilities.
  - Use *Kernel* functions as a mappings of similarity measures in a transformed space where nonlinearities can be solved linearly.
  - Two widely used kernel functions are:

$$K_\sigma(\bar{x}, \bar{z}) = \exp\Big(-\sigma \sum_k (x_k - z_k)^2\Big) \tag{1}$$

and

$$K_d(\bar{x}, \bar{z}) = \langle \bar{x}, \bar{z} \rangle^d, \tag{2}$$

Problem and Motivation

Example

Problem and Motivation

Example

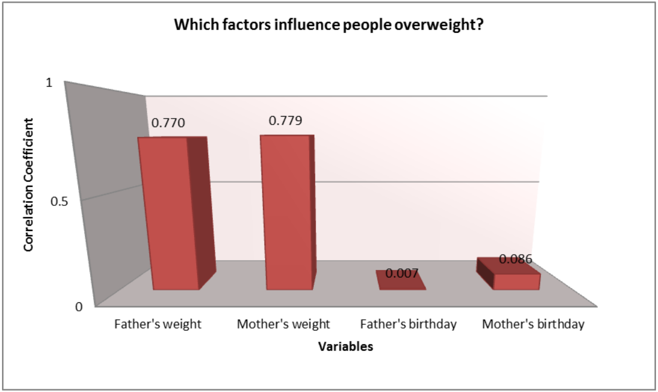


Figure: Correlation analysis of input variables.

Problem and Motivation

Example

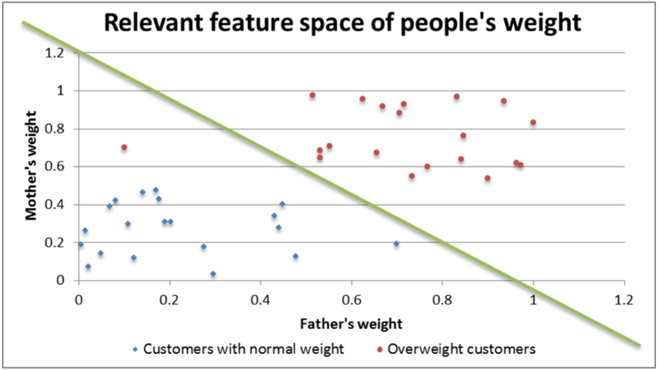


Figure: Relevant variables helps to build better predictor models.

Problem and Motivation

- Some FSS techniques (Filters) assume variables are independent and are very popular because of low computational cost.
- Real world behaves different because variables may influence others but this interaction is usually hidden i.e. in Bioinformatic.
- Prediction accuracy could be affected if dependencies are ignored.
- Multivariate FSS methods search for feature subsets and possible dependency relationships between them.
- The challenge is then to design novel feature selection methods that take advantage of multivariate power combined with high-accuracy classifiers, such as kernel classifiers, to obtain improved prediction and explanatory performance.

Research Hypothesis

Research Hypothesis

Given a set of observations taken from a particular phenomenon where dependencies among variables exist but are hidden, the selection of a subset of relevant variables is *feasible* with a method that combines iterative estimation of dependency networks coupled with weighted-kernel classifiers, in such a way the method will provide a predictor model that *improves* the understanding of the problem domain when compared with other techniques based on independence assumptions.

Background

wKIERA (weigthed kernel iterative estimation of relevance algorithm.)

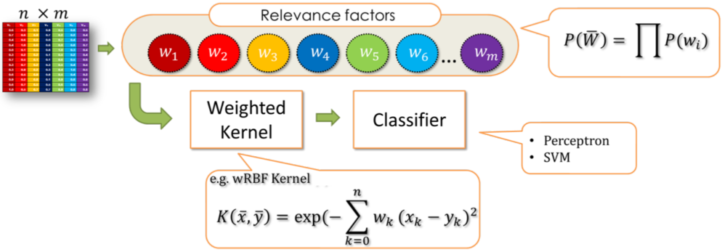


Figure: Main components of the wKIERA algorithm.

Background

FSS-EBNA (Feature subset selection by estimation of Bayesian network algorithm.)

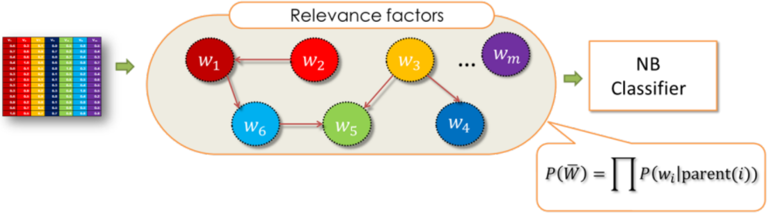


Figure: Main components of the FSS-EBNA algorithm.

Idea and Proposal

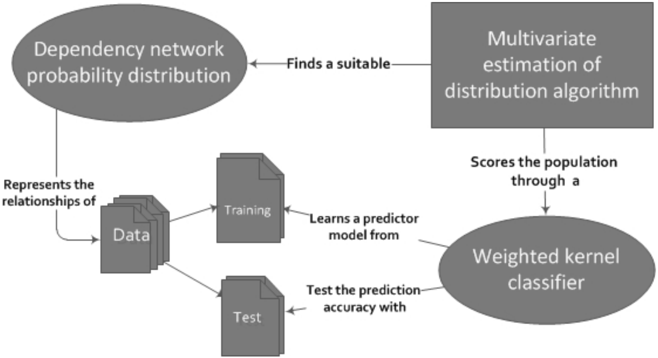


Figure: The components involved in this thesis proposal.

Idea and Proposal

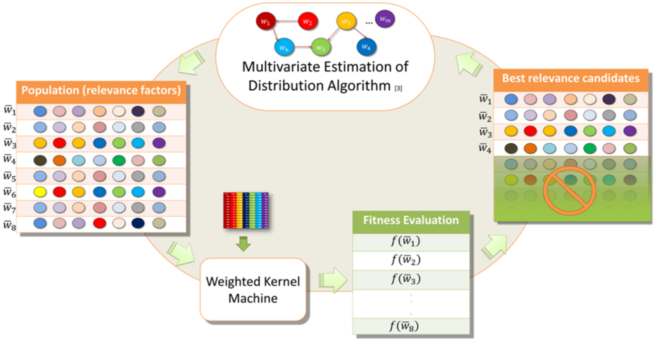


Figure: Preliminary depiction of the expected method described in this proposal.

Goals

Goals

- *Main*
  Design an algorithm for feature relevance estimation by estimating dependency networks combined with weighted kernel classification methods.

Goals

- *Main*
  Design an algorithm for feature relevance estimation by estimating dependency networks combined with weighted kernel classification methods.
- *Secondary*

Goals

- *Main*
  Design an algorithm for feature relevance estimation by estimating dependency networks combined with weighted kernel classification methods.
- *Secondary*
  - Verify the feasibility of the algorithm in high dimensional feature spaces using toy and real datasets (e.g bioinformatics).

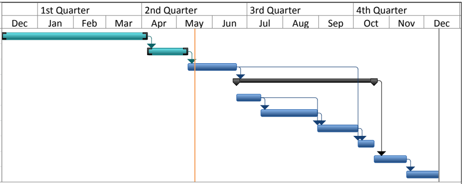
Goals

- *Main*

  Design an algorithm for feature relevance estimation by estimating dependency networks combined with weighted kernel classification methods.
- *Secondary*
  - Verify the feasibility of the algorithm in high dimensional feature spaces using toy and real datasets (e.g bioinformatics).
  - Compare algorithm performance with respect to other feature selection methods (e.g. score-based filters, stochastic population-based wrappers and weighted-kernel-based embedded methods).

Timeline

ID	Task Name
1	Literature Review
2	Algorithm Design
3	Algorithm Implementation
4	**Experimentation**
5	Data set preparation
6	Setup comparison algorithms
7	Experimental design
8	Experiments execution
9	Analysis of Results
10	Conclusions and Documentation



Project: Time line
Date: Tue 5/17/11

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

Algorithm Design

Algorithm 1 Preliminary pseudocode of the expected method described in this proposal

Inputs: Given a dataset $\mathcal{D}$, a weighted kernel κ_ω and a classifier $\mathcal{A}$

Let β represents a dependency network distribution initialized with an independent joint distribution: $\beta \leftarrow$ Independent joint distribution.

repeat

  Split $\mathcal{D}$ in training $\mathcal{D}_\alpha$ and testing $\mathcal{D}_\theta$ data

  $\bar{\Omega} \leftarrow$ Sample k candidates from β

  for $\boldsymbol{\omega}_j \in \bar{\Omega}$ **do**

    Train classifier: $h_j \leftarrow \mathcal{A}(\mathcal{D}_\alpha, \kappa_{\omega j})$

    Test classifier: $s_j \leftarrow \mathrm{error}(h_j, \mathcal{D}_\theta, \kappa_{\omega j})$

  end for

  $\bar{\Omega}' \leftarrow \mathrm{bestCandidates}(\bar{\Omega}, s)$

  Re-estimate dependency network: $\beta \leftarrow \mathrm{reEstimate}(\bar{\Omega}')$

until Dependency network has converge or maximum iterations reached

References

[1] Shumeet Baluja and R. Caruana. Removing the genetics from the standard genetic algorithm. In *The Proceedings of the 12th Annual Conference on Machine Learning*, pages 38 – 46. Morgan Kaufmann Publishers, 1995.

[2] Shummet Baluja. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical report, Carnegie Mellon University Pittsburgh, PA, USA, Pittsburgh, PA, USA, 1994.

[3] Endika Bengoetxea, Pedro Larrañaga, Isabelle Bloch, and Aymeric Perchant. Estimation of distribution algorithms: A new evolutionary computation approach for graph matching problems. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 2134 of *Lecture Notes in Computer Science*, pages 454–469. Springer Berlin / Heidelberg, 2001.

[4] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002. 10.1023/A:1012450327387.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 10.1007/BF00994018.

[6] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press, 1 edition, March 2000.

[7] Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296, December 1999.

[8] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[9] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Machine Learning*, 3:1157–1182, March 2003.

[10] Georges Harik, Fernando G. Lobo, and David E. Goldberg. The compact genetic algorithm. In *IEEE Transactions on Evolutionary Computation*, pages 523–528, 1998.

[11] David Heckerman. *A tutorial on learning with Bayesian networks*, pages 301–354. MIT Press, Cambridge, MA, USA, 1999.

[12] Mary S. Lipton and Ljiljana Pasa-Tolic. *Mass Spectrometry of Proteins and Peptides: Methods and Protocols*. Humman Press, 2008.

[13] H. Muhlenbein and G. Paag. From recombination of genes to the estimation of distributions: I. binary parameters. In Hans-Michael Voigt, Werner Ebeling, Ingo Rechenberg, and Hans-Paul Schwefel, editors, *Parallel Problem Solving from Nature PPSN IV*, volume 1141 of *Lecture Notes in Computer Science*, pages 178–187. Springer Berlin / Heidelberg, 1996.

[14] Heinz Muhlenbein, Thilo Mahnig, and Alberto Ochoa Rodriguez. Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics*, 5:215–247, July 1999.

[15] Pedro Larra naga, Ramon Etxeberria, Jose A. Lozano, and Jose M. Pe na. Combinatorial optimization by learning and simulation of bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 343–352. Morgan Kaufmann, 2000.

[16] Iñaki Inza, Marisa Merino, Pedro Larra naga, Jorge Quiroga, Basilio Sierra, and Marcos Girala. Feature subset selection by genetic algorithms and estimation of distribution algorithms. a case study in the survival of cirrhotic patients treated with tips. *BioData Mining*, 2000.

[17] Iñaki Inza, Pedro Larra naga, Ramon Etxeberria, and Basilio Sierra. Feature subset selection by bayesian network-based optimization. *Artificial Intelligence*, pages 157–184, 2000.

[18] Rubn Arma nanzas, Iñaki Inza, Roberto Santana, Yvan Saeys, Jose Luis Flores, Jose Antonio Lozano, Yves Van De Peer, Rosa Blanco, Vctor Robles, Concha Bielza, and Pedro Larra naga. A review of estimation of distribution algorithms in bioinformatics. *BioData Mining*, 2008.

[19] Martin Pelikan, David E. Goldberg, and Erick Cantu-Paz. Boa: The bayesian optimization algorithm. In *Genetic and Evolutionary Computation Conference (GECCO-1999)*, pages 525–532. Morgan Kaufmann, 1999.

[20] Sergio Rojas and Delmiro Fernandez-Reyes. Adapting multiple kernel parameters for support vector machines using genetic algorithms. In *2005 IEEE Congress on Evolutionary Computation (CEC-2005)*, 2005.

[21] Sergio Rojas, Emily Hsieh, Dan Agranoff, Sanjeev Krishna, and Delmiro Fernandez-Reyes. Estimation of relevant variables on high-dimensional biological patterns using iterated weighted kernel functions. *PLoS ONE*, 3:e1806, 03 2008.

[22] Frank Rosenblatt. *The perceptron: a probabilistic model for information storage and organization in the brain*, pages 386–408. American Psychological Association, 1956.

[23] Yvan Saeys, Sven Degroeve, Dirk Aeyels, and Yves Van De Peer. Fast feature selection using a simple estimation of distribution algorithm: A case study on splice site prediction. *Bioinformatics*, 19, 2003.

[24] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, September 2007.

[25] Kumara Sastry, David E. Goldberg, and Xavier Llora. Towards billion bit optimization via parallel estimation of distribution algorithm. In *Genetic and Evolutionary Computation Conference (GECCO-2007)*, pages 577–584, 2007.

[26] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.