The University of Manchester
Manchester Business School

MANCHESTER 1824

UNIVERSITEIT GENT

Generalizing Random Forests Principles to other Methods: Random MultiNomial Logit, Random Naive Bayes, …

Anita Prinzie & Dirk Van den Poel

UNIVERSITEIT GENT

Random Forests

- Forest of decision trees
- Introduced by Breiman (2001) to reduce instability of decision tree:

" ... unstable learning algorithms- algorithms whose output classifier undergoes major changes in response to small changes in the training data. Decision-tree, neural network, and rule learning algorithms are all unstable."

Dietterich (2000)

UNIVERSITEIT GENT

Prediction Error

= Irreducible error + statistical bias + variance
Hastie, Tibshirani and Friedman, Springer, 2001, p. 37.

Improve predictive performance by reducing bias and variance
Dietterich and Kong (1995)

Decision Tree and Prediction Error

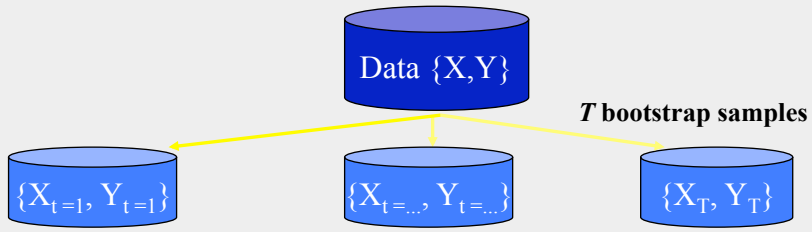
Low statistical bias but moderate to high variance
Dietterich and Kong (1995)

"Bagging can dramatically reduce the variance of unstable procedures like trees, leading to improved prediction."

Hastie, Tibshirani and Friedman, Springer, 2001, p. 247.

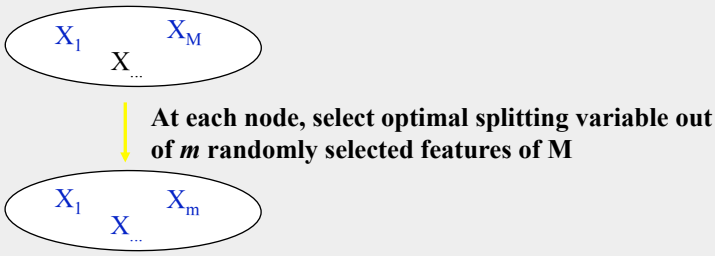
Random Forests (Breiman 2001) and Randomness

Random observations: bagging (Breiman, 1996)



Random Forests and Randomness

Random features: Increase noise robustness (Breiman, 2001)

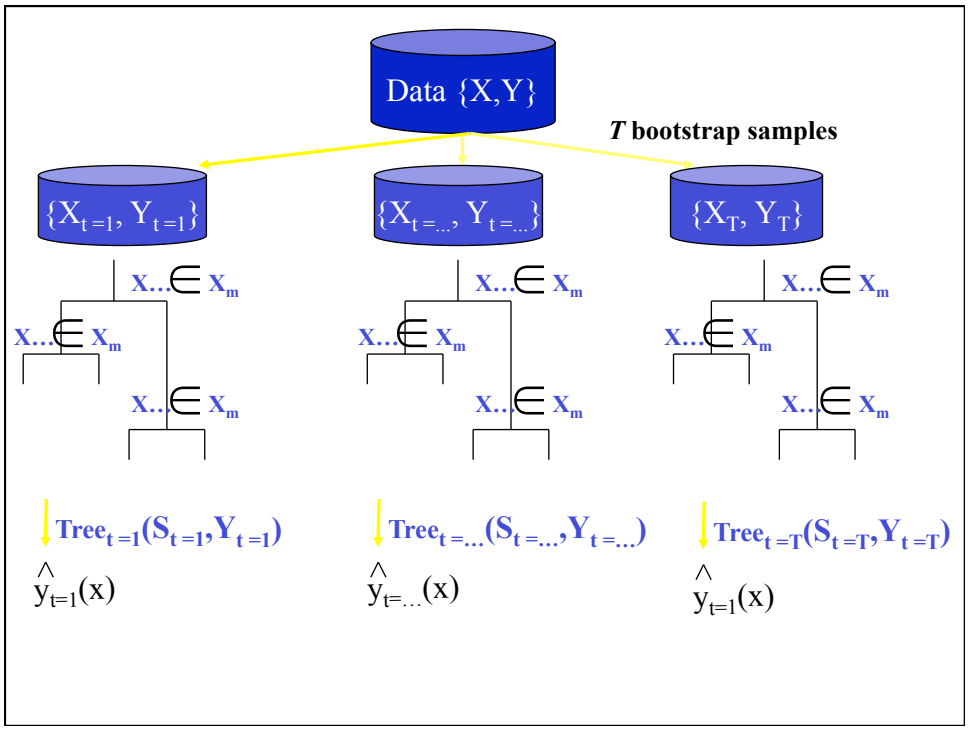


Noise Robustness

Robust to many input variables with each one containing only a small amount of information

Breiman (2001)

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be



RF and Classification

Put input vector down the *T* trees in the forest.

Each tree votes for the predicted class.

Majority voting:

Classify instance into class having the most votes over all trees *T* in the forest

RF and Feature Importance

Importance of unique feature *m*

a) Calculate performance on instances left out of *t*-th decision tree (Out-Of-Bag)

b) Randomly permute *m*-th feature in the OOB data and apply respective *t*-th decision tree

a- b, average over all decision trees containing feature *m* and standardize

Research Question

Could generalizing the Random Forests framework to other methods improve their predictive performance?

	RF	METHOD
bagging	Stability of decision tree => reduced variance	Stability of method Ensemble and accuracy (Dietterich, 2000)
random feature	Noise robustness	Hughes phenomenon => better accuracy Noise robustness

Hughes Phenomenon (Hughes, 1968)

The density of a fixed amount of data points *decreases* as the number of dimensions *increases*, inhibiting the ability to make reliable estimates of the probability distribution

- optimal number of features for given number of observations
- on increasing the number of features as input to a method over a given threshold, the accuracy decreases

Random Utility Models and MNL

Multinomial logit dominant RU model due to closed-form solution

MNL and multicollinearity

- ⇒ MNL not well-suited for large feature spaces
- ⇒ Need for feature selection in MNL

	MNL	RF
Stability	Yes	Yes
Theory	Yes	No
Large feature space	No	**Yes**
Feature importance	Biased	**Unbiased**

	MNL	RMNL	
Stability	Yes		
Theory	Yes		
Large feature space			**Yes**
Feature importance			**Unbiased**

	RMNL	RNB
bagging	Stability of even stable classifier Ensemble and accuracy (Dietterich, 2000)	Ensemble and accuracy (Dietterich, 2000) Cf. AODE (Webb, Boughton, & Wang, 2005)
random feature	Multicollinearity => better accuracy Noise robustness	Alleviate conditional independence: subsets Cf. SBC (Langley and Sage, 1994)

Random MultiNomial Logit

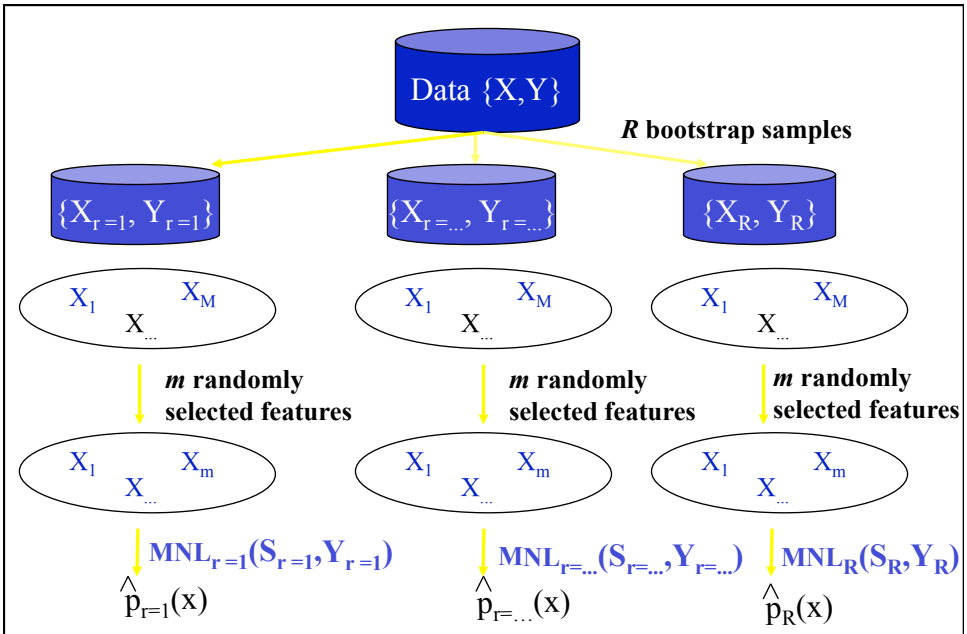
RF => RMNL

1. Random Forests

   ⇒ Forest of *T* decision trees selecting optimal split at each node out of *m* randomly selected features of M, estimated on *T* bootstrap samples S_T

2. RMNL

   ⇒ Forest of *R* MNLs with *m* randomly selected features out of *M*, estimated on *R* bootstrap samples S_R



Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

RMNL and Classification

Deliver input vector to *R* MNLs and infer predicted class by adjusted Majority Voting

$$\hat{p}_{RMNL} = \frac{1}{R}\sum_{r=1}^{R}\hat{p}_{r_{MNL}}$$

$$\hat{y} = \arg\max \sum_{k=1}^{K}\hat{p}_{k_{RMNL}}$$

Random MNL

RMNL's importance of feature *m*

a) Calculate performance on instances left out of *r*-th MNL (Out-Of-Bag)

b) Randomly permute *m*-th feature in the OOB data, apply respective *r*-th MNL, calculate peformance

a-b, average over all MNLs containing feature *m* and standardize

Random Naive Bayes

Naive Bayes

Class-conditional independence:

Effect of variable value on class is independent of the values of other variables

Biased posterior probabilities but … accurate and fast!

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

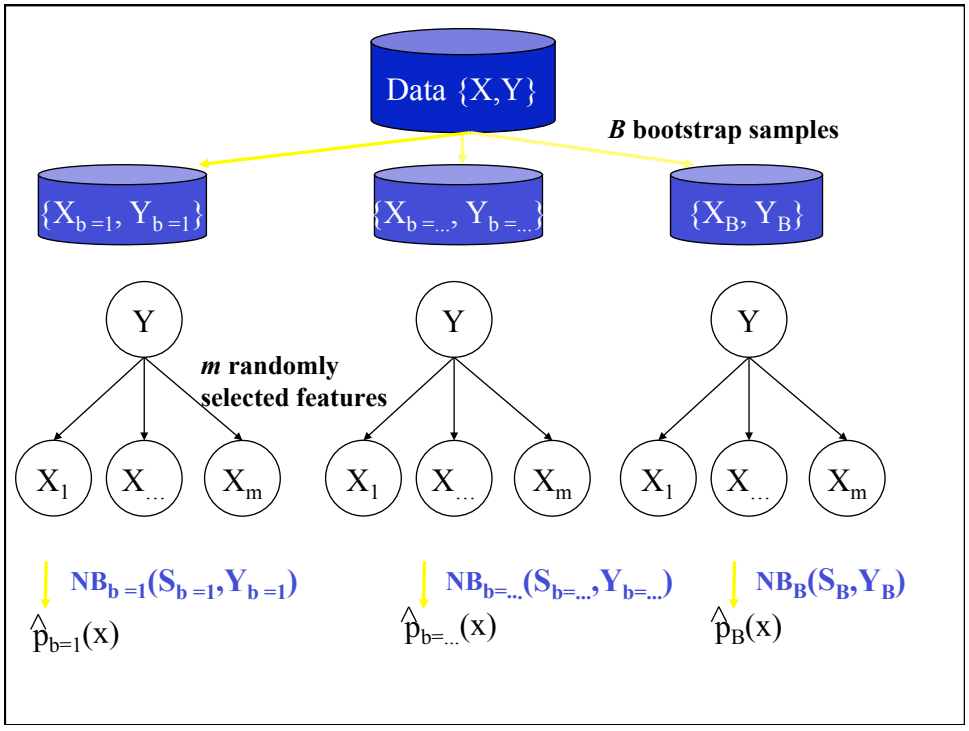
	RMNL	RNB
bagging	Stability of even stable classifier Ensemble and accuracy (Dietterich, 2000)	Ensemble and accuracy (Dietterich, 2000) Cf. AODE (Webb, Boughton, & Wang, 2005)
random feature	Multicollinearity => better accuracy Noise robustness	Alleviate conditional independence: subsets Cf. SBC (Langley and Sage, 1994)

Random Naive Bayes

RF => RNB

Forest of *B* NBs with *m* randomly selected features out of *M*, estimated on *B* bootstrap samples S_B

Class-conditional independence within subset of *m* randomly chosen variables out of feature space *M*



UNIVERSITEIT GENT

Random Naive Bayes

RNB

Classification: adjusted Majority Voting

$$\hat{p}_{RNB} = \frac{1}{R}\sum_{r=1}^{R} \hat{p}_{r_{NB}} \qquad \hat{y} = \arg\max \sum_{k=1}^{K} \hat{p}_{k_{RNB}}$$

Feature importances on oob data

UNIVERSITEIT GENT

Application ...
Customer Intelligence

Application

NPTB Model for CROSS-SELL Analysis

Predict in what product category a customer will buy next ⇕ product

Scanner data from Belgian home-appliance retailer. N_train= 37,276
N_test = 37,110

Application

Cleaning, communication, entertainment and cooking needs

➡ 9 product categories

	Product Category	Products	Need
1	**Washing / Drying**	washing machine, clothes dryer, dish washer	Cleaning clothes and dishes
2	**House Cleaning**	vacuum cleaner, carpet cleaner	Cleaning house
3	**Mobiles**	mobile phones, telephone, fax, answering machine	Communication
4	**VCR / DVD**	VCR, DVD, projector	Entertainment: audio+visual (luxury)
5	**TV**		Entertainment: visual
6	**Audio**	hi-fi, equalizer, amplifier	Entertainment: audio
7	**Fun Cooking**	toaster, food processor	Cooking: preparation of food
8	**Cooking**	refrigerator, freezer, range	Cooking: non portable
9	**Cooking Small**	microwave oven, deep fryer, oven	Cooking: portable

NULL

Predictors X:

- Monetary value, purchase depth & width
- # Home appliances acquired at retailer
- Socio-demo
- Brand loyalty
- Price sensitivity
- # Home-appliances returned
- Dominant mode of payment
- Special life-event experience
- ORDER of acquisition of appliances
- Time to first-acquisition or repeated acquisition (DURATION)

➡ 89 dimensions

More details can be found in:

PRINZIE, A., & VAN DEN POEL, D. (2007), Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models, *Decision Support Systems,* 44 (1), 28-45.

Predictive Model Evaluation

- Test set instead of oob data
- wPCC and AUC

$$w_k = \frac{1 - f_k}{\sum_{k=1}^{K} 1 - f_k} \quad \text{s.t.} \quad \sum_{k=1}^{K} w_k = 1$$

$$wPCC = \sum_{k=1}^{K} wPCC_k \quad \text{with} \quad wPCC_k = w_k \times PCC_k$$

Evaluation

	wPCCt	PCCt	AUCt
RF			
MNL			
RMNL			
NB			
RNB			
SVM			

Estimation

RF on all M=441 features

- *T*=500
- Balanced RF
- Optimize number of random features *m* to split each node on. Default m: $\sqrt{M}$=21, stepsize 1/3 => [7,399] ⇒ *m**=336

MNL and Expert Feature Selection

- MNL with *M*=89 did not converge
- Expert Feature Selection (wrapper) on NULL, Order and Duration features

(Prinzie and Van den Poel, Decision Support Systems, 44, 2007, 28-45)

MNL Expert Feature Selection

Model	wPCCe
BEST3_5	0.1884
BEST3_5+ORDER	0.1923
Best3_5+DURATION	0.1984
BEST3_5+ORDER+DURATION	0.1956

$$> Cr_{pro} = \sum_{k=1}^{K} f_k^2 = 0.1228$$

Morrison, JMR (1969)

Estimation

RMNL on all M=89 features

- R=100
- Optimize number of random features m to estimate R MNLs on. Default m: $\sqrt{M}$=9, stepsize 1/3 => [3,84] ⇨

m^*=48

Estimation

Combining fit Members

- All 100 MNLs with fixed m value have performance better than Cr_{pro}

•"Combining only very accuracte classifiers might improve the performance of the ensemble even more."
Dietterich (1997)

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

Estimation

RMNL combining MNLs with 10% highest wPCC for given *m*

- *R*=10
- Optimize number of random features *m* to estimate *R* MNLs on. Default *m*: √*M*=9, stepsize 1/3 => [3,84]

*m**=48

Estimation

NB

- *Laplace estimation, M=441*
- *Fayad and Irani (1993) supervised discretisation*

Estimation

RNB on all M=441 features

- B=100
- Optimize number of random features m to estimate B NBs on. Default m: $\sqrt{M}$=21, stepsize 1/3 => [7,399] ⇒ m^*=42

Estimation

RNB combining NBs with 10% highest wPCC, M=441

- B=10
- Optimize number of random features m to estimate B MNLs on. Default m: $\sqrt{M}$=21, stepsize 1/3 => [7,399] ⇒ m^*=77

Estimation

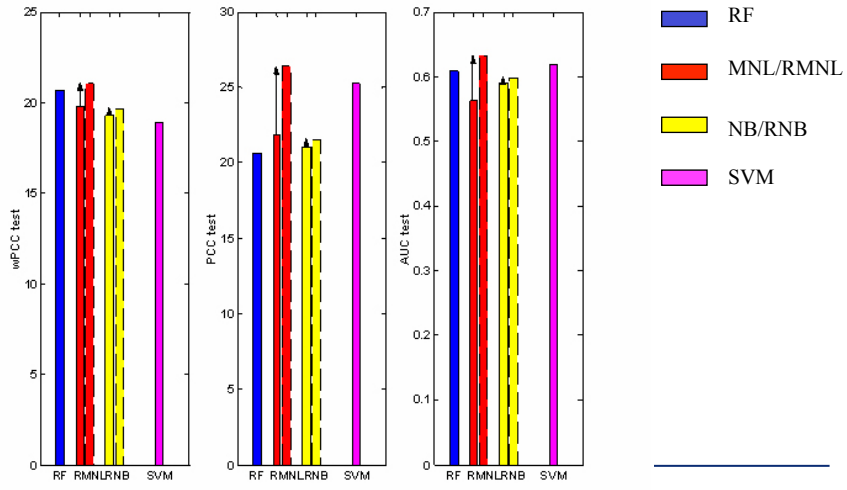
SVM

- *Libsvm is used (Chang & Lin, 2001)*
- Optimal (C,ɣ) is determined by grid search using five-fold cross validation

Validation

	wPCCt	PCCt	AUCt
RF	20.66	21.39	0.6090
MNL	19.87	21.84	0.5926
RMNL_10	**21.06**	**26.41**	**0.6322**
NB	19.27	21.05	0.5899
RNB_10	19.61	21.56	0.5983
SVM	18.92	25.24	0.6188

Validation



Validation

Statistical significance of differences in *k* class-specific AUCs

- Non-parametric test (Delong, Delong and Clarke-Pearson,1998)
- RMNL_10 versus RF, MNL, NB, RNB_10 and SVM: all significant at α=0.05 except for RMNL_10-SVM, k=4 and k=5.

Statistical significance of differences in *k* class-specific AUCs

- RNB_10 versus NB: all *k* AUCs are statistically significant different at α=0.05.

Top-20 Features in NPTB Model

Rank	Varname	z	Block	Description
1	productnbr_pc	29.37	1	monetary, depth and width
2	diffproduct_pc	24.91	1	monetary, depth and width
3	gender	19.70	3	socio-demo
4	ORDER Markov 2nd order	16.01	9	order
5	DURATION (surv)	9.48	10	duration
6	ORDER Markov 2nd order	9.21	9	order
7	ORDER dummies	7.69	9	order
8	ORDER Markov for Discrimination	4.86	9	order
9	language	4.84	3	socio-demo
10	nbrdiffbrand	4.74	4	brand loyalty
11	loyal_PANASONIC	4.51	4	brand loyalty
12	ORDER Markov 2nd order	4.44	9	order
13	rnbrreturns	4.41	6	returns
14	nbrabovep90	4.32	5	price sensitivity
15	maxdiffprod	3.96	2	number acquired
16	nbrbelowq1	3.87	5	price sensitivity
17	maxprod	3.74	2	number acquired
18	maxamount	3.38	1	monetary, depth and width
19	DURATION (survdiff)	3.36	10	duration
20	ORDER Markov 2nd order	3.34	9	order

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

Conclusion

Generalization of Random Forests to other methods is valuable!

Introduction of Random MultiNomial Logit and Random Naive Bayes as two possible generalizations of Random Forests

Random MNL

RMNL accommodates 2 weaknesses of MNL:

1) The curse of dimensionality
2) The sensitivity to multicollinearity and as a consequence the biased feature estimates

and

1) improves accuracy
2) provides feature importances

Random Naive Bayes

1) Alleviates the class-conditional independence assumption
2) Improves accuracy
3) Provides feature importances

Random Forests, Random MNL and Random Naive Bayes:

1) Random observation selection
2) Random feature selection
3) Tuning parameter: *m*
4) No need for test/validation data set
5) Generalization error and feature importance estimates on oob data

Future Research

- Generalize Random Forests to other methods potentially benefiting from random observation selection and random feature selection:

  supervised: classification or regression
  unsupervised: cluster ensembles

Future Research

- Ensemble with flexible *m* across base methods
- Multi-target problems

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

Software

Independent runs
=> easily to parallellize

Given results, priority is given to RMNL

Acknowledgement

Leo Breiman

Symposium on Data Mining, May 10th, 2004, Gent, Belgium.

Anita Prinzie & Dirk Van den Poel – www.crm.ugent.be

PRINZIE, A., & VAN DEN POEL, D.,
Random Forests for multiclass classification: Random MultiNomial Logit, *Expert Systems with Applications*, 34(3), 2008, 1721-1732.

DOI:
10.1016/j.eswa.2007.01.029

PRINZIE, A., & VAN DEN POEL, D.,
Random Multiclass Classification: Generalizing RF to RMNL and RNB, *Dexa 2007, Lecture Notes in Computer Science, 4653, 349-358.*

http://www.springerlink.com/content/6l71084745234516/

The University of Manchester
Manchester Business School

MANCHESTER 1824

UNIVERSITEIT GENT

Questions, suggestions, ...

www.crm.UGent.be
www.mma.UGent.be

anita.prinzie@mbs.ac.uk,UGent.be
dirk.vandenpoel@UGent.be

UNIVERSITEIT GENT