

# Ensemble Learning: Bayesian Model Averaging

Fabio Cimmino

Università degli Studi di Milano-Bicocca

Anno Accademico 2017-2018

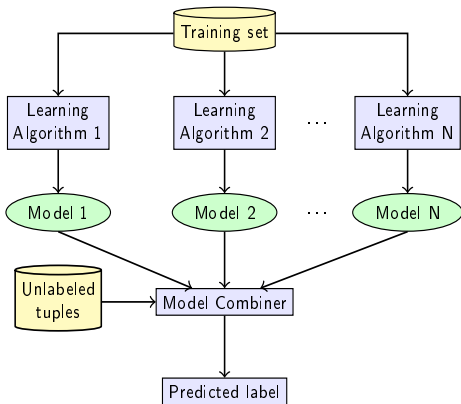


Relatore: Prof.ssa Vincenzina Messina  
Co-Relatore: Dott.ssa Elisabetta Fersini



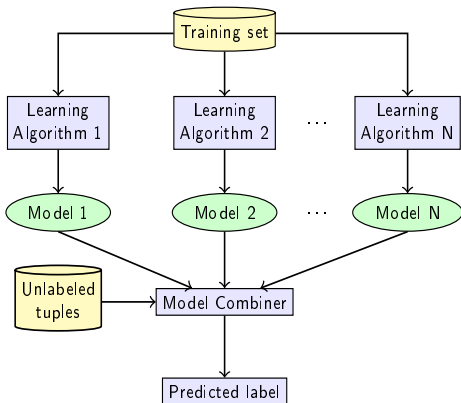
# Ensemble Learning

L'Ensemble Learning consiste in metodi d'insieme che usano **modelli multipli** per ottenere una miglior prestazione predittiva **rispetto ai singoli modelli**



# Ensemble Learning

L'Ensemble Learning consiste in metodi d'insieme che usano **modelli multipli** per ottenere una miglior prestazione predittiva **rispetto ai singoli modelli**

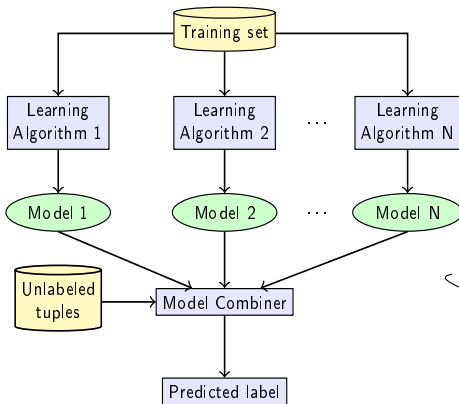


Vantaggio principale Ensemble Learning:

- Diminuisce il limite decisionale che separa i dati dalle diverse classi eliminando l'incertezza sul modello da usare

# Ensemble Learning

L'Ensemble Learning consiste in metodi d'insieme che usano **modelli multipli** per ottenere una miglior prestazione predittiva **rispetto ai singoli modelli**



Vantaggio principale Ensemble Learning:

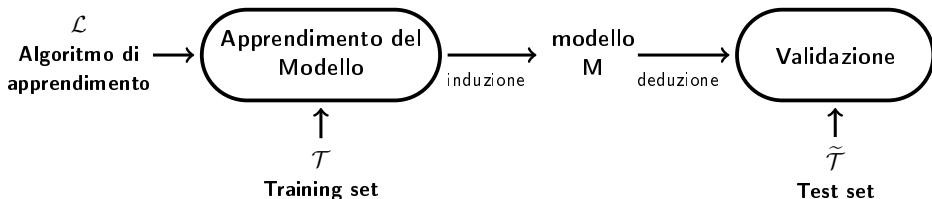
- Diminuisce il limite decisionale che separa i dati dalle diverse classi eliminando l'incertezza sul modello da usare

Attributi

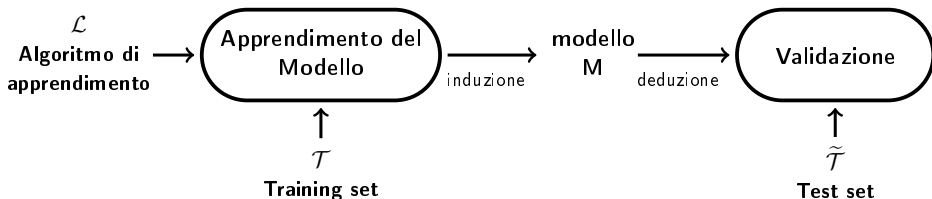
Classe

$A_1$	$A_2$	...	$A_m$	$C$
$t_1[A_1]$	$t_1[A_2]$	...	$t_1[A_m]$	$t_1[C]$
$\vdots$				$\vdots$
$t_n[A_1]$	$t_n[A_2]$	...	$t_n[A_m]$	$t_n[C]$

# Validazione del modello e performance

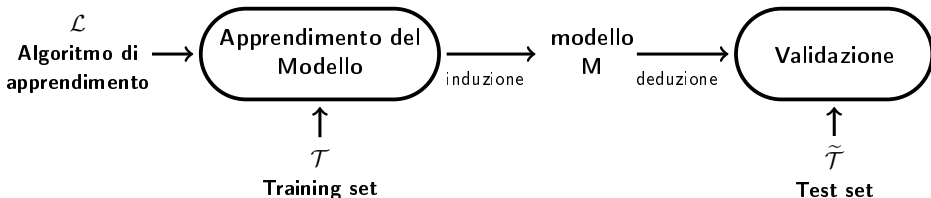


# Validazione del modello e performance



		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

# Validazione del modello e performance



		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

$$Accuracy = \frac{True\ Positive + True\ Negative}{\sum Total\ Population}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F\ measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

# Stato dell'arte e limitazioni

- Bagging
- Boosting
- Stacking
- Random Forests



# Stato dell'arte e limitazioni

- Bagging
- Boosting
- Stacking
- Random Forests

Limitazioni:

- 1 Non vengono considerate le capacità di generalizzazione dei modelli

# Stato dell'arte e limitazioni

- Bagging
- Boosting
- Stacking
- Random Forests

## Limitazioni:

- 1 Non vengono considerate le capacità di generalizzazione dei modelli
- 2 Modelli indipendenti e ugualmente affidabili

# Stato dell'arte e limitazioni

- Bagging
- Boosting
- Stacking
- Random Forests

## Limitazioni:

- 1 Non vengono considerate le capacità di generalizzazione dei modelli
- 2 Modelli indipendenti e ugualmente affidabili
- 3 La ricerca dell'insieme più accurato ha un costo

# Bayesian Model Averaging

$l^*(r) \equiv$  classe finale assegnata ad un record del dataset

$D \equiv$  dataset

$S \equiv$  possibile insieme di modelli  $S \subseteq C$

$$\begin{aligned}
 l^*(r) = \arg \max P(l(r) \mid S, D) &= \sum_{i \in S} P(l(r) \mid i, D) P(i \mid D) \\
 &= \sum_{i \in S} P(l(r) \mid i, D) P(D \mid i) P(i) \\
 &= \sum_{i \in S} P(l(r) \mid i, D) P(D \mid i)
 \end{aligned}$$

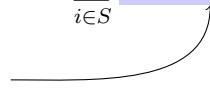
# Bayesian Model Averaging

$l^*(r) \equiv$  classe finale assegnata ad un record del dataset

$D \equiv$  dataset

$S \equiv$  possibile insieme di modelli  $S \subseteq C$

$$\begin{aligned}
 l^*(r) = \arg \max P(l(r) \mid S, D) &= \sum_{i \in S} P(l(r) \mid i, D) P(i \mid D) \\
 &= \sum_{i \in S} P(l(r) \mid i, D) P(D \mid i) P(i) \\
 &= \sum_{i \in S} \underbrace{P(l(r) \mid i, D)}_{\text{blue}} \underbrace{P(D \mid i)}_{\text{red}}
 \end{aligned}$$

- Probabilità marginale di  $l(r)$  



# Bayesian Model Averaging

$l^*(r) \equiv$  classe finale assegnata ad un record del dataset

$D \equiv$  dataset

$S \equiv$  possibile insieme di modelli  $S \subseteq C$

$$\begin{aligned}
 l^*(r) = \arg \max P(l(r) \mid S, D) &= \sum_{i \in S} P(l(r) \mid i, D) P(i \mid D) \\
 &= \sum_{i \in S} P(l(r) \mid i, D) P(D \mid i) P(i) \\
 &= \sum_{i \in S} \boxed{P(l(r) \mid i, D)} \boxed{P(D \mid i)}
 \end{aligned}$$

- Probabilità marginale di  $l(r)$  
- Capacità di generalizzazione del modello  $i$  



# Bayesian Model Averaging

$l^*(r) \equiv$  classe finale assegnata ad un record del dataset

$D \equiv$  dataset

$S \equiv$  possibile insieme di modelli  $S \subseteq C$

$$\begin{aligned}
 l^*(r) = \arg \max P(l(r) \mid S, D) &= \sum_{i \in S} P(l(r) \mid i, D) P(i \mid D) \\
 &= \sum_{i \in S} P(l(r) \mid i, D) P(D \mid i) P(i) \\
 &= \sum_{i \in S} \boxed{P(l(r) \mid i, D)} \boxed{P(D \mid i)}
 \end{aligned}$$

- Probabilità marginale di  $l(r)$  
- Capacità di generalizzazione del modello  $i$  

$P(D \mid i) \approx$  F-measure del training set

# Selezione dell'insieme ottimale

Contributo  $r_i^s$  di ogni modello  $i$  appartenente ad un determinato insieme  $S \subseteq C$ :

$$r_i^s = \frac{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)}$$



# Selezione dell'insieme ottimale

Contributo  $r_i^s$  di ogni modello  $i$  appartenente ad un determinato insieme  $S \subseteq C$ :

$$r_i^s = \frac{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)}$$

composizione ottimale  
con *backward elimination* :  $\frac{ACC(S)}{|S|} \geq \frac{ACC(S \setminus x)}{|S - 1|}$

# Selezione dell'insieme ottimale

Contributo  $r_i^s$  di ogni modello  $i$  appartenente ad un determinato insieme  $S \subseteq C$ :

$$r_i^s = \frac{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)}$$

composizione ottimale  
con *backward elimination* :  $\frac{ACC(S)}{|S|} \geq \frac{ACC(S \setminus x)}{|S| - 1}$

$\sum_{p=1}^N \frac{N!}{p!(N-p)!}$  possibili soluzioni

- $N$  modelli totali
- Insieme ottimale composto da  $p$  modelli

# Selezione dell'insieme ottimale

Contributo  $r_i^s$  di ogni modello  $i$  appartenente ad un determinato insieme  $S \subseteq C$ :

$$r_i^s = \frac{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)}$$

composizione ottimale  
con *backward elimination* :  $\frac{ACC(S)}{|S|} \geq \frac{ACC(S \setminus x)}{|S| - 1}$

$\sum_{p=1}^N \frac{N!}{p!(N-p)!}$  possibili soluzioni



$N - 1$  potenziali modelli candidati

- $N$  modelli totali
- Insieme ottimale composto da  $p$  modelli

# Algoritmi ed insiemi di modelli utilizzati

## Algoritmi di apprendimento:

- K-Nearest Neighbors
- Decision Tree
- Multilayer Perceptron
- Naive Bayes
- Support Vector Machines

# Algoritmi ed insiemi di modelli utilizzati

## Algoritmi di apprendimento:

- K-Nearest Neighbors
- Decision Tree
- Multilayer Perceptron
- Naive Bayes
- Support Vector Machines

6 insiemi di modelli:

5 omogenei

1 eterogeneo

- 10 modelli di K-Nearest Neighbors
- 10 modelli di Decision Tree
- 10 modelli di Multilayer Perceptron
- 3 modelli di Naive Bayes
- 40 modelli di Support Vector Machines

- 5 modelli prodotti dai 5 algoritmi

# Dataset

Dataset	# Istanze	# Attributi	# Classi
anneal	898	39	5
autos	205	26	6
audiology	226	70	2
balance-scale	625	5	3
breast-cancer	286	10	2
breast-w	699	10	2
colic	368	23	2
credit-rating	690	16	2
german-credit	1000	21	2
pima-diabetes	768	9	2
glass	214	10	6
heart-c	303	14	2
heart-h	294	14	5
heart-statlog	270	14	2
hepatitis	155	20	2
hypothyroid	3772	30	4

Dataset	# Istanze	# Attributi	# Classi
ionosphere	351	35	2
iris	150	5	3
kr-vs-kp	3196	37	2
labor	57	17	2
lymph	148	19	4
mushroom	8124	23	2
primary-tumor	339	18	9
segment	2310	20	7
sick	3772	30	2
sonar	208	61	2
soybean	683	36	19
vehicle	846	19	4
vote	435	17	2
vowel	990	13	11
zoo	101	18	7

# Risultati Insieme Omogeneo Multilayer Perceptron

Dataset	Boosting	Bagging	Stacking	RandomForest	Democratic	Bayesian
anneal	0.8363	0.9822	0.7617	0.9933	<u>0.9933</u>	<u>0.9933</u>
autos	0.4488	0.6976	0.3268	<u>0.8341</u>	0.7938	0.7986
audiology	0.4646	0.7655	0.2522	0.7699	<u>0.8401</u>	0.8314
balance-scale	0.7272	0.8288	0.4576	0.8048	<u>0.9247</u>	0.9152
breast-cancer	0.7028	0.6783	0.7028	0.6923	<u>0.7452</u>	0.7384
breast-w	0.9485	0.9557	0.6552	0.9614	0.9642	0.9685
colic	0.8125	0.8533	0.6404	<u>0.8614</u>	0.8559	<u>0.8478</u>
credit-rating	0.8464	0.8507	0.5505	<u>0.8507</u>	0.8696	<u>0.8696</u>
german-credit	0.695	0.744	0.70	0.725	<u>0.7710</u>	0.7690
pima-diabetes	0.7435	0.7461	0.651	0.7383	0.7683	<u>0.7722</u>
glass	0.4486	0.6963	0.3551	<u>0.729</u>	0.7190	<u>0.7100</u>
heart-c	0.8218	0.8218	0.5446	0.8152	0.8483	<u>0.8548</u>
heart-h	0.7789	0.7857	0.6395	0.7789	<u>0.8575</u>	<u>0.8508</u>
heart-statlog	0.8	0.7926	0.5556	0.7815	<u>0.8407</u>	<u>0.8407</u>
hepatitis	0.8258	<u>0.8452</u>	0.7935	0.8258	0.8396	0.8329
hypothyroid	0.9321	<u>0.9955</u>	0.9229	0.991	0.9510	0.9531
ionosphere	0.9088	0.9088	0.641	0.9288	0.9317	0.9203
iris	0.9533	0.94	0.3333	0.9533	<u>0.9800</u>	<u>0.9800</u>
kr-vs-kp	0.9384	0.9912	0.5222	0.9881	<u>0.9950</u>	0.9947
labor	0.8772	0.8596	0.6491	0.8772	0.8967	<u>0.9333</u>
lymph	0.7432	0.7838	0.5473	0.8108	<u>0.8519</u>	<u>0.8519</u>
mushroom	0.962	<u>1</u>	0.518	<u>1</u>	<u>1</u>	<u>1</u>
primary-tumor	0.2891	0.4513	0.2478	0.4248	<u>0.4840</u>	0.4721
segment	0.2857	0.9697	0.1429	<u>0.9766</u>	0.9680	0.9684
sick	0.9719	<u>0.9849</u>	0.9388	0.9838	0.9751	0.9761
sonar	0.7163	0.774	0.5337	0.8077	<u>0.8374</u>	0.8326
soybean	0.2796	0.8682	0.1318	0.9165	0.9458	<u>0.9487</u>
vehicle	0.3995	0.727	0.2565	0.7707	<u>0.8583</u>	0.8570
vote	0.954	0.9586	0.6138	0.9586	<u>0.9609</u>	<u>0.9679</u>
vowel	0.1737	0.8576	0.909	0.9606	0.9636	<u>0.9677</u>
zoo	0.604	0.4257	0.4059	0.8911	<u>0.9618</u>	<u>0.9618</u>

# Risultati Insieme Omogeneo Multilayer Perceptron

Il 45% delle volte il Bayesian Model Averaging risulta la tecnica migliore con l'insieme omogeneo Multilayer Perceptron

Dataset	Boosting	Bagging	Stacking	RandomForest	Democratic	Bayesian
anneal	0.8363	0.9822	0.7617	0.9933	<u>0.9933</u>	<u>0.9933</u>
autos	0.4488	0.6976	0.3268	<u>0.8341</u>	0.7938	0.7986
audiology	0.4646	0.7655	0.2522	0.7699	<u>0.8401</u>	0.8314
balance-scale	0.7272	0.8288	0.4576	0.8048	<u>0.9247</u>	0.9152
breast-cancer	0.7028	0.6783	0.7028	0.6923	<u>0.7452</u>	0.7384
breast-w	0.9485	0.9557	0.6552	0.9614	0.9642	0.9685
colic	0.8125	0.8533	0.6404	<u>0.8614</u>	0.8559	<u>0.8478</u>
credit-rating	0.8464	0.8507	0.5505	<u>0.8507</u>	0.8696	<u>0.8696</u>
german-credit	0.695	0.744	0.70	0.725	<u>0.7710</u>	0.7690
pima-diabetes	0.7435	0.7461	0.651	0.7383	0.7683	<u>0.7722</u>
glass	0.4486	0.6963	0.3551	<u>0.729</u>	0.7190	<u>0.7100</u>
heart-c	0.8218	0.8218	0.5446	0.8152	0.8483	<u>0.8548</u>
heart-h	0.7789	0.7857	0.6395	0.7789	<u>0.8575</u>	<u>0.8508</u>
heart-statlog	0.8	0.7926	0.5556	0.7815	<u>0.8407</u>	<u>0.8407</u>
hepatitis	0.8258	<u>0.8452</u>	0.7935	0.8258	0.8396	0.8329
hypothyroid	0.9321	<u>0.9955</u>	0.9229	0.991	0.9510	0.9531
ionosphere	0.9088	0.9088	0.641	0.9288	0.9317	0.9203
iris	0.9533	0.94	0.3333	0.9533	<u>0.9800</u>	<u>0.9800</u>
kr-vs-kp	0.9384	0.9912	0.5222	0.9881	<u>0.9950</u>	0.9947
labor	0.8772	0.8596	0.6491	0.8772	0.8967	<u>0.9333</u>
lymph	0.7432	0.7838	0.5473	0.8108	<u>0.8519</u>	<u>0.8519</u>
mushroom	0.962	<u>1</u>	0.518	<u>1</u>	<u>1</u>	<u>1</u>
primary-tumor	0.2891	0.4513	0.2478	0.4248	<u>0.4840</u>	0.4721
segment	0.2857	0.9697	0.1429	<u>0.9766</u>	0.9680	0.9684
sick	0.9719	<u>0.9849</u>	0.9388	0.9838	0.9751	0.9761
sonar	0.7163	0.774	0.5337	0.8077	<u>0.8374</u>	0.8326
soybean	0.2796	0.8682	0.1318	0.9165	0.9458	<u>0.9487</u>
vehicle	0.3995	0.727	0.2565	0.7707	<u>0.8583</u>	0.8570
vote	0.954	0.9586	0.6138	0.9586	<u>0.9609</u>	<u>0.9679</u>
vowel	0.1737	0.8576	0.909	0.9606	0.9636	<u>0.9677</u>
zoo	0.604	0.4257	0.4059	0.8911	<u>0.9618</u>	<u>0.9618</u>



# Risultati Insieme Eterogeneo

Dataset	Boosting	Bagging	Stacking	RandomForest	Democratic	Bayesian
anneal	0.8363	0.9822	0.7617	<u>0.9933</u>	0.9922	0.9922
autos	0.4488	0.6976	0.3268	<u>0.8341</u>	0.8181	<u>0.8529</u>
audiology	0.4646	0.7655	0.2522	0.7699	<u>0.8449</u>	0.8225
balance-scale	0.7272	0.8288	0.4576	0.8048	<u>0.9087</u>	0.9024
breast-cancer	0.7028	0.6783	0.7028	0.6923	<u>0.7590</u>	0.7558
breast-w	0.9485	0.9557	0.6552	0.9614	0.9699	0.9714
colic	0.8125	0.8533	0.6404	<u>0.8614</u>	0.8559	<u>0.8586</u>
credit-rating	0.8464	0.8507	0.5505	<u>0.8507</u>	<u>0.8638</u>	<u>0.8638</u>
german-credit	0.695	0.744	0.70	0.725	<u>0.7690</u>	0.7640
pima-diabetes	0.7435	0.7461	0.651	0.7383	<u>0.7800</u>	0.7800
glass	0.4486	0.6963	0.3551	0.729	<u>0.7201</u>	<u>0.7472</u>
heart-c	0.8218	0.8218	0.5446	0.8152	0.8347	<u>0.8415</u>
heart-h	0.7789	0.7857	0.6395	0.7789	<u>0.8540</u>	<u>0.8506</u>
heart-statlog	0.80	0.7926	0.5556	0.7815	<u>0.8556</u>	0.8593
hepatitis	0.8258	0.8452	0.7935	0.8258	0.8508	<u>0.8638</u>
hypothyroid	0.9321	0.9955	0.9229	0.991	0.9690	<u>0.9960</u>
ionosphere	0.9088	0.9088	0.641	0.9288	0.9517	0.9460
iris	0.9533	0.94	0.3333	0.9533	<u>0.9800</u>	<u>0.9800</u>
kr-vs-kp	0.9384	0.9912	0.5222	0.9881	0.9947	<u>0.9956</u>
labor	0.8772	0.8596	0.6491	0.8772	<u>0.9333</u>	<u>0.9333</u>
lymph	0.7432	0.7838	0.5473	0.8108	<u>0.8710</u>	0.8581
mushroom	0.962	<u>1</u>	0.518	<u>1</u>	<u>1</u>	<u>1</u>
primary-tumor	0.2891	0.4513	0.2478	0.4248	0.4807	<u>0.4837</u>
segment	0.2857	0.9697	0.1429	0.9766	0.9801	<u>0.9805</u>
sick	0.9719	0.9849	0.9388	0.9838	0.9812	<u>0.9852</u>
sonar	0.7163	0.774	0.5337	0.8077	<u>0.8702</u>	0.8657
soybean	0.2796	0.8682	0.1318	0.9165	0.9458	<u>0.9473</u>
vehicle	0.3995	0.727	0.2565	0.7707	0.8217	<u>0.8264</u>
vote	0.954	0.9586	0.6138	0.9586	<u>0.9701</u>	0.9678
vowel	0.1737	0.8576	0.909	0.9606	0.9707	<u>0.9929</u>
zoo	0.604	0.4257	0.4059	0.8911	0.9718	<u>0.9809</u>

# Risultati Insieme Eterogeneo

Il 65% delle volte il Bayesian Model Averaging risulta la tecnica migliore con l'insieme eterogeneo

Dataset	Boosting	Bagging	Stacking	RandomForest	Democratic	Bayesian
anneal	0.8363	0.9822	0.7617	0.9933	0.9922	0.9922
autos	0.4488	0.6976	0.3268	0.8341	0.8181	0.8529
audiology	0.4646	0.7655	0.2522	0.7699	0.8449	0.8225
balance-scale	0.7272	0.8288	0.4576	0.8048	0.9087	0.9024
breast-cancer	0.7028	0.6783	0.7028	0.6923	0.7590	0.7558
breast-w	0.9485	0.9557	0.6552	0.9614	0.9699	0.9714
colic	0.8125	0.8533	0.6404	0.8614	0.8559	0.8586
credit-rating	0.8464	0.8507	0.5505	0.8507	0.8638	0.8638
german-credit	0.695	0.744	0.70	0.725	0.7690	0.7640
pima-diabetes	0.7435	0.7461	0.651	0.7383	0.7800	0.7800
glass	0.4486	0.6963	0.3551	0.729	0.7201	0.7472
heart-c	0.8218	0.8218	0.5446	0.8152	0.8347	0.8415
heart-h	0.7789	0.7857	0.6395	0.7789	0.8540	0.8506
heart-statlog	0.80	0.7926	0.5556	0.7815	0.8556	0.8593
hepatitis	0.8258	0.8452	0.7935	0.8258	0.8508	0.8638
hypothyroid	0.9321	0.9955	0.9229	0.991	0.9690	0.9960
ionosphere	0.9088	0.9088	0.641	0.9288	0.9517	0.9460
iris	0.9533	0.94	0.3333	0.9533	0.9800	0.9800
kr-vs-kp	0.9384	0.9912	0.5222	0.9881	0.9947	0.9956
labor	0.8772	0.8596	0.6491	0.8772	0.9333	0.9333
lymph	0.7432	0.7838	0.5473	0.8108	0.8710	0.8581
mushroom	0.962	1	0.518	1	1	1
primary-tumor	0.2891	0.4513	0.2478	0.4248	0.4807	0.4837
segment	0.2857	0.9697	0.1429	0.9766	0.9801	0.9805
sick	0.9719	0.9849	0.9388	0.9838	0.9812	0.9852
sonar	0.7163	0.774	0.5337	0.8077	0.8702	0.8657
soybean	0.2796	0.8682	0.1318	0.9165	0.9458	0.9473
vehicle	0.3995	0.727	0.2565	0.7707	0.8217	0.8264
vote	0.954	0.9586	0.6138	0.9586	0.9701	0.9678
vowel	0.1737	0.8576	0.909	0.9606	0.9707	0.9929
zoo	0.604	0.4257	0.4059	0.8911	0.9718	0.9809

# Conclusioni e sviluppi futuri

Il successo di un sistema d'insieme si basa direttamente sulla **diversità** degli algoritmi di apprendimento che compongono l'insieme

# Conclusioni e sviluppi futuri

Il successo di un sistema d'insieme si basa direttamente sulla **diversità** degli algoritmi di apprendimento che compongono l'insieme

Sviluppi futuri:

- 1 Sviluppo di una strategia di *forward selection*

# Conclusioni e sviluppi futuri

Il successo di un sistema d'insieme si basa direttamente sulla **diversità** degli algoritmi di apprendimento che compongono l'insieme

Sviluppi futuri:

- 1 Sviluppo di una strategia di *forward selection*
- 2 Parallelizzazione del Bayesian Model Averaging

# Conclusioni e sviluppi futuri

Il successo di un sistema d'insieme si basa direttamente sulla **diversità** degli algoritmi di apprendimento che compongono l'insieme

Sviluppi futuri:

- 1 Sviluppo di una strategia di *forward selection*
- 2 Parallelizzazione del Bayesian Model Averaging
- 3 Sviluppo di una struttura di *multi-task learning*