Ensembles of selected classifiers and clusters

B Lausen, K Stoyanov, H Nordmark, A Perperoglou

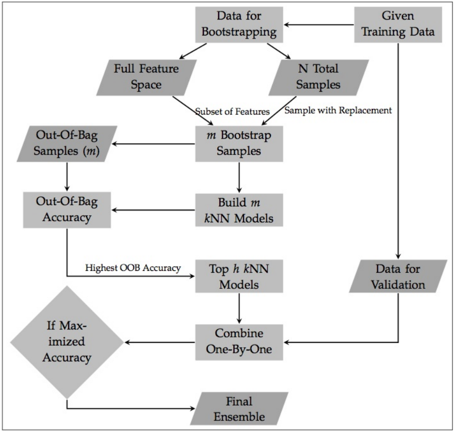
based on joint work on new methods with W Adler, PO Degens, A Gul, T Hothorn, Z Khan, O Mahmoud, R Medellin, S Potapov and M Schumacher

Overview

1. Ensembles of selected classifiers
2. Optimal selection of cutpoints
3. Ensemble methods for cluster analysis

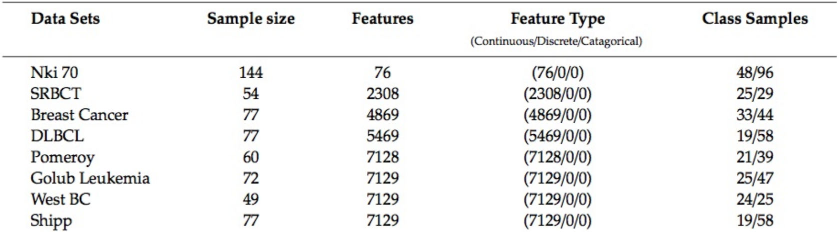
Ensemble of Subset of *k*-Nearest Neighbours Models

Gul et al. (ECDA2014, Springer 2016); R CRAN package ESKNN (2015)



Ensemble of Subset of *k*-Nearest Neighbours Models

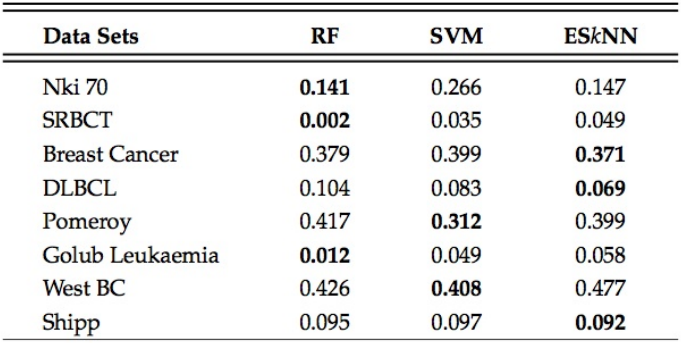
Bench mark data sets



Data Sets	Sample size	Features	Feature Type (Continuous/Discrete/Catagorical)	Class Samples
Nki 70	144	76	(76/0/0)	48/96
SRBCT	54	2308	(2308/0/0)	25/29
Breast Cancer	77	4869	(4869/0/0)	33/44
DLBCL	77	5469	(5469/0/0)	19/58
Pomeroy	60	7128	(7128/0/0)	21/39
Golub Leukemia	72	7129	(7129/0/0)	25/47
West BC	49	7129	(7129/0/0)	24/25
Shipp	77	7129	(7129/0/0)	19/58

Ensemble of Subset of *k*-Nearest Neighbours Models

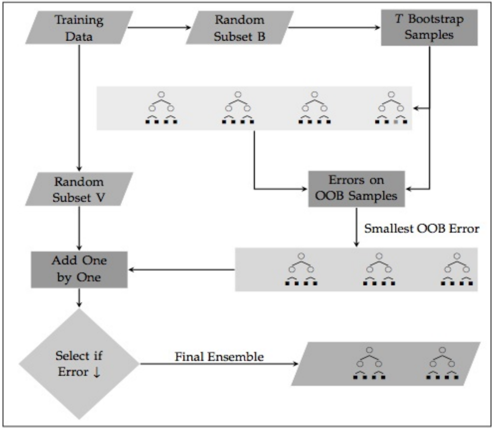
Results



Data Sets	RF	SVM	ES*k*NN
Nki 70	**0.141**	0.266	0.147
SRBCT	**0.002**	0.035	0.049
Breast Cancer	0.379	0.399	**0.371**
DLBCL	0.104	0.083	**0.069**
Pomeroy	0.417	**0.312**	0.399
Golub Leukaemia	**0.012**	0.049	0.058
West BC	0.426	**0.408**	0.477
Shipp	0.095	0.097	**0.092**

Ensemble of Optimal Trees

Khan et al. (ECDA2014, Springer 2016); R CRAN package OTE (2015)



Data Set	*n*	*d*	FT	kNN	Tree	RF	NH	SVM	SVM	SVM	SVM	OTE
			(R/I/N)					(Radial)	(Linear)	(Bessel)	(Lapla.)	
Mammogr.	830	5	(0/5/0)	0.141	0.123	0.129	**0.121**	0.134	0.125	0.131	0.135	0.137
Dystrophy	209	5	(2/3/0)	0.105	0.134	0.095	0.116	0.083	0.087	0.080	**0.079**	0.086
Monk3	122	6	(0/6/0)	0.089	0.069	0.066	0.182	0.070	0.157	0.066	0.094	**0.061**
Appendicitis	106	7	(6/0/0)	0.126	0.135	0.120	0.117	0.1360	0.1257	**0.116**	0.118	0.124
SAHeart	462	9	(5/3/1)	0.209	0.207	0.190	0.188	0.185	**0.179**	0.197	0.182	0.201
tic-tac-toe	958	9	(0/0/9)	0.228	0.147	**0.041**	0.120	0.148	0.219	0.120	0.197	0.044
Heart	303	13	(1/12/0)	0.222	0.168	**0.123**	0.144	0.144	0.128	0.123	0.125	0.129
House vote	232	16	(0/0/16)	0.066	0.032	0.029	0.066	0.030	0.035	0.158	0.039	**0.029**
Bands	365	19	(13/6/0)	0.223	0.255	0.188	0.224	0.199	0.203	0.223	0.211	**0.181**
Hepatitis	80	20	(2/18/0)	0.311	0.138	0.097	0.095	0.096	0.104	0.116	0.089	**0.088**
Parkinson	195	22	(22/0/0)	0.115	0.114	0.068	0.093	0.076	0.120	0.154	0.093	**0.064**
Body	507	23	(22/1/0)	0.019	0.073	0.031	0.055	0.012	**0.012**	0.238	0.022	0.030
Thyroid	9172	27	(3/2/22)	0.031	0.010	0.008	0.016	0.039	0.032	0.057	0.038	**0.008**
WDBC	569	29	(29/0/0)	0.054	0.064	0.031	0.043	0.027	**0.021**	0.203	0.028	0.031
WPBC	198	32	(30/2/0)	0.183	0.213	0.168	0.169	0.160	**0.154**	0.181	0.163	0.165
Oil-Spill	937	49	(40/9/0)	0.040	0.033	0.028	0.029	0.033	0.037	0.033	0.036	**0.027**
Spam base	4601	57	(55/2/0)	0.174	0.095	0.038	0.091	0.073	0.062	0.241	0.081	**0.037**
Glaucoma	196	62	(62/0/0)	0.137	0.110	**0.089**	0.092	0.094	0.124	0.219	0.119	0.090
Nki 70	144	76	(71/5/0)	0.146	0.141	0.147	0.147	0.168	0.202	0.235	0.183	**0.133**
Musk	476	166	(0/166/0)	0.142	0.188	0.096	0.175	0.096	0.110	0.247	0.189	**0.087**

Fair method: *p*-value adjusted classification trees

We prespecify a value $p_{stop} > 0$ and a value or rule n_{min} for the minimum size of the leafs of the tree. We start with the set consisting of all observations:

Fair method: *p*-value adjusted classification trees

We prespecify a value $p_{stop} > 0$ and a value or rule n_{min} for the minimum size of the leafs of the tree. We start with the set consisting of all observations:

a) The adjusted minimal *p*-value p_k is computed for all K factors and all allowable splits within the factors.

Fair method: *p*-value adjusted classification trees

We prespecify a value $p_{stop} > 0$ and a value or rule n_{min} for the minimum size of the leafs of the tree. We start with the set consisting of all observations:

a) The adjusted minimal *p*-value p_k is computed for all K factors and all allowable splits within the factors.

b) The set of objects is split into two subsets based on the factor $\hat{k}$ and the corresponding cutpoint $\hat{\mu}$, if the adjusted *p*-value for K factors is smaller or equal to the prespecified value p_{stop}:

$$P_{H_0}(M(\mathbf{a}, \mathbf{X}, \varepsilon_1, \varepsilon_2) > b) \leq p_{stop} \ .$$

Fair method: *p*-value adjusted classification trees

We prespecify a value $p_{stop} > 0$ and a value or rule n_{min} for the minimum size of the leafs of the tree. We start with the set consisting of all observations:

a) The adjusted minimal *p*-value p_k is computed for all K factors and all allowable splits within the factors.

b) The set of objects is split into two subsets based on the factor $\hat{k}$ and the corresponding cutpoint $\hat{\mu}$, if the adjusted *p*-value for K factors is smaller or equal to the prespecified value p_{stop}:

$$P_{H_0}(M(\mathbf{a}, \mathbf{X}, \varepsilon_1, \varepsilon_2) > b) \leq p_{stop} \; .$$

c) The partition procedure is stopped, if there exists no allowable split, or if $P_{H_0}(M(\mathbf{a}, \mathbf{X}, \varepsilon_1, \varepsilon_2) > b) > p_{stop}$ or because of the n_{min} criterion.

Fair method: *p*-value adjusted classification trees

We prespecify a value $p_{stop} > 0$ and a value or rule n_{min} for the minimum size of the leafs of the tree. We start with the set consisting of all observations:

a) The adjusted minimal *p*-value p_k is computed for all K factors and all allowable splits within the factors.

b) The set of objects is split into two subsets based on the factor $\hat{k}$ and the corresponding cutpoint $\hat{\mu}$, if the adjusted *p*-value for K factors is smaller or equal to the prespecified value p_{stop}:

$$P_{H_0}(M(\mathbf{a}, \mathbf{X}, \varepsilon_1, \varepsilon_2) > b) \leq p_{stop} \ .$$

c) The partition procedure is stopped, if there exists no allowable split, or if $P_{H_0}(M(\mathbf{a}, \mathbf{X}, \varepsilon_1, \varepsilon_2) > b) > p_{stop}$ or because of the n_{min} criterion.

d) For each of the two subsets we repeat this procedure.

Clustering methods:

- Idea following Chen et al (2005) proposal 'Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis'
- Improvement of algorithm; idea to estimate the number of clusters by hierarchical clustering
- R package work in progress

Cluster evaluation:

- Nonparametric bootstrap evaluation (Felsenstein 1985)
- Parametric bootstrap evaluation (Lausen and Degens 1988)
- Statistical model $d = d_u + e$; three-objects-variance estimation can be seen as an ensemble method; ...

Discussion

- Combination of POS and Ensemble of Subset of k-Nearest Neighbours Models demonstrate possible gains
- OTE may allow further gains
- Improvement of gene selection using Proportional Overlapping Score (POS)
- Features measured on different scales - p-value adjusted classification trees
- Ensemble methods and evaluation for cluster analysis
- (Ensemble) cluster methods for feature selection to improve predictive accuracy

References

- GUL, A., PERPEROGLOU, A., KHAN, Z., MAHMOUD, O., MIFTAHUDDIN, M., ADLER, W., Lausen, B. (2016), Ensemble of a subset of kNN classifiers, *Advances in Data Analysis and Classification, (online first)*.
- KHAN, Z., GUL, A., MAHMOUND, O., MIFTAHUDDIN, M., PERPEROGLOU, A., ADLER, W., LAUSEN, B. (2016), An Ensemble of Optimal Trees for Class Membership Probability Estimation, In: Wilhelm, A., Kestler, H. A. (eds.), *Analysis of Large and Complex Data*, Springer-Verlag Berlin.
- LAUSEN, B., DEGENS, P.O. (1988), Evaluation of the reconstruction of phylogenies with DNA-DNA hybridization data, in: Bock, H.H. (ed.), *Classification and related methods of data analysis*, Proceedings First conference of the international federation of classification societies (IFCS), North Holland, Amsterdam, 367-374.
- MAHMOUD, O., HARRISON, A.P., PERPEROGLOU, A., GUL, A., KHAN, Z., METODIEV, M., LAUSEN, B. (2014), A feature selection method for classification within functional genomics experiments based on the proportional overlapping score, *BMC Bioinformatics, 15 (1). p. 274*.
- STOYANOV, K. (2015), Hierarchical k-means clustering and its application in customer segmentation. Master dissertation, Department of Mathematical Sciences, University of Essex, UK.