Gul, A., Khan, Z., Perperoglou, A., Mahmoud, O., Miftahuddin, M., Adler, W., & Lausen, B. (2016). Ensemble of subset of k-nearest neighbours models for class membership probability estimation. In *Analysis of Large and Complex Data* (pp. 411-421). (Studies in Classification, Data Analysis, and Knowledge Organization). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-319-25226-1_35

Ensemble of Subset of *k*-Nearest Neighbours Models for Class Membership Probability Estimation

Asma Gul[1], Zardad Khan[1], Aris Perperoglou[1], Osama Mahmoud[1], Miftahuddin Miftahuddin[1], Werner Adler[2], and Berthold Lausen[1]

[1] Department of Mathematical Sciences, University of Essex, Colchester, UK.
agul@essex.ac.uk
[2] Department of Biometry and Epidemiology, University of Erlangen-Nuremberg, Germany

Abstract. Combining multiple classifiers can give substantial improvement in prediction performance of learning algorithms especially in the presence of non-informative features in the data sets. This technique can also be used for estimating class membership probabilities. We propose an ensemble of *k* Nearest Neighbours (*k*NN) classifiers for class membership probability estimation in the presence of non-informative features in the data. This is done in two steps. Firstly, we select classifiers based upon their individual performance from a set of base *k*NN models, each generated on a bootstrap sample using a random feature set from the feature space of training data. Secondly, a step wise selection is used on the selected learners, and those models are added to the ensemble that maximize its predictive performance. We use bench mark data sets with some added non-informative features for the evaluation of our method. Experimental comparison of the proposed method with usual *k*NN, bagged *k*NN, random *k*NN and random forest shows that it leads to high predictive performance in terms of minimum Brier score on most of the data sets. The results are also verified by simulation studies.

Keywords

ENSEMBLE METHODS, *k*-NEAREST NEIGHBOURS, NON-INFORMATIVE FEATURES

1 Introduction

In numerous real life applications, class membership probabilities of individuals are required in addition to their class labels. For example, in safety-critical domains such as surgery, oncology, internal medicine, pathology, paediatrics and human genetics, these probabilities are needed. In all the aforementioned areas, probability estimates are more useful than simple classification as they provide a measure of reliability of the decision taken about an individual (Lee et al. (2010), Malley et al. (2012), Kruppa et al. (2012), Kruppa et al. (2014a, 2014b)). Machine learning techniques used mainly for classification can be

used as non-parametric methods for class membership probability estimation in order to avoid the assumptions imposed in parametric models used for the estimation of these probabilities (Kruppa et al. (2012), Malley et al. (2012)).

In many real life problems, one often encounters imprecise data such as data with non-informative features. These features dramatically decrease the prediction performance of the algorithms (Nettleton et al. (2010)). Feature selection methods that investigate the most discriminative features from the original features are usually recommended to mitigate the effect of such non-informative features (Mahmoud et al. 2014a, 2014b). However, different feature selection methods result in different feature subsets for the same data set thus varying feature relevancy. This encourages combining the results of several best feature subsets.

It has been investigated in the last two decades that combining the outputs of multiple models, known as ensemble techniques, results in improved prediction performance (Breiman (1996), Hothorn and Lausen (2003), Kuncheva (2004)) and are more resilient to non-informative features in the data than using an individual model (Melville et al. (2004)). Recently, an ensemble of optimal trees has been suggested for class membership probability estimation by Khan et al. (2015).

k Nearest Neighbour (kNN) learning algorithm is one of the simplest and oldest methods. It classifies an unknown observation to the class of majority among its k nearest neighbour points in the training data as measured by a distance metric (Cover and Hart (1967)). Despite its simplicity, kNN gives competitive results and in some cases even outperforms other complex learning algorithms. However, kNN is vulnerable to non-informative features in the data. Attempts have been made by researchers to improve the performance of Nearest Neighbour algorithm by ensemble techniques (Bay (1998), Li et al. (2011), Samworth (2012)). In this manuscript, we propose an ensemble of subset of kNN classifiers (ESkNN) for the task of estimating class membership probability, particularly in the presence of non-informative features in the data set and compare the results with those of simple kNN, bagged kNN (BkNN), random kNN (RkNN) and random forest (RF).

2 Proposed Ensemble of Subset of kNN algorithm

To construct the ensemble of subset of kNN models (ESkNN), a two stage strategy is implemented. Consider a training data set of $n\times(d+1)$ dimensions, consisting of data points $\mathcal{L} = (\mathbf{x}, \mathbf{y})$, an instance is characterized by d features along with the corresponding class label. The training data set $\mathcal{L}$ is randomly divided into two sets, a learning set and validation set and the ensemble is developed in the following steps.

1. Draw m random feature sets of size l from d input features, $l < d$ and draw m bootstrap samples on these feature sets from the learning set.

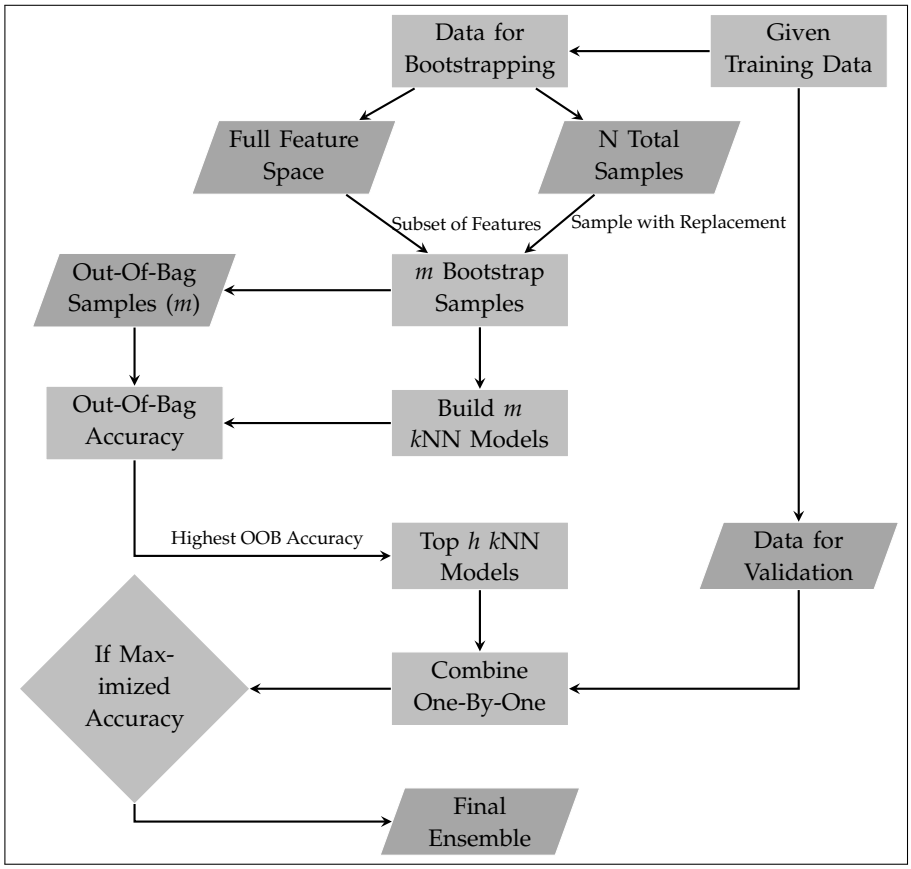


Fig. 1. A flow chart of the steps of *ESkNN* for class membership probability estimation.

2. Build m kNN models and select h of the m models that give the highest accuracy on the out-of-bag observations (observations that are left out from the bootstrap sample).
3. In the next stage add the selected h models one by one starting from the best model and assess its collective performance on the validation set. The process is repeated until all the h models are evaluated in the ensemble.
4. A model is selected if it gives minimum Brier score (BS) on the validation set. Let $BS^{\langle r-1\rangle}$ be the Brier score of the ensemble without the rth model and $BS^{\langle r\rangle}$ is the Brier score after adding the model, the rth model is selected if
$$BS^{\langle r\rangle} < BS^{\langle r-1\rangle}. \tag{1}$$
5. The group membership probability estimate of the test instance is the averaged probability estimate over all t selected models.

A flow chart of the procedure of ESkNN is shown in Figure 1.

3 Performance Measure of the Methods

As a performance measure, we use Brier score introduced by Brier (1950). It provides a measure of accuracy of the predicted probabilities. It is the most common and appropriate criterion for binary class outcome and can be used for the evaluation of predicted probabilities by a machine learning algorithm, in situations where the true probabilities are unknown (Malley et al. (2012)). Gneiting and Raftery (2007) stated that the Brier score is a proper measure and its minimum value can only occur if the calculated probabilities are taken as the true probabilities which are unknown. It follows that a machine learning technique that has the smallest value of the Brier score will be performing best in estimating group membership probabilities. The Brier score is given as:

$$BS = E(y_i - p(y_i|\mathbf{x}))^{\mathbf{2}}, \tag{2}$$

where $y_i \in \{0, 1\}$ and $p(y_i|\mathbf{x})$ is the true but unknown probability of the state of the outcome for y_i given the features. An estimator for the above score for t test observations is:

$$\hat{BS} = \frac{\sum_{i=1}^{\mathrm{t}} \left(y_i - \hat{p}(y_i|\mathbf{x})\right)^2}{\mathrm{t}}. \tag{3}$$

4 Results and Discussion

4.1 Experiments and Discussion on Bench Mark Problems

The performance of the ESkNN in terms of the Brier score, is evaluated on a total of twenty five data sets taken from UCI, KEEL databases and from within R-Libraries, mlbench, mboost, ipred, gclus and mmst. Summary of the data sets is given in table 1.

The ESkNN is assessed in two scenarios. Firstly, all the methods are applied on the data sets with their original features and secondly, the feature space of all the data sets are extended by adding 500 randomly generated non-informative features. The results for both the cases are given in Table 2 and 3. Each of the data sets is divided into test and training parts where the training part consists of 90% of observations and the remaining 10% is reserved for testing. The methods are applied on each data set in a total of 1000 runs and are evaluated on the testing data set in each run. The final Brier score is the average of Brier scores of the 1000 runs. The experiments are carried out using the R-Program. The values of the hyper parameters for the methods are selected by using the "tune" function within the R-Package "e1071".

The results from Table 2 reveal that ESkNN is giving the smallest Brier scores on 23 data sets out of the total 25 data sets among all kNN based methods, whereas on Bands data set it gives better probability estimate than kNN and BkNN and comparable to RkNN. When comparing to random forest it gives low Brier scores on 10 data sets.

Table 1. Summary of the data sets. The first five data sets are microarray.

Data Sets	Sample size	Features	Feature Type (Continuous/Discrete/Catagorical)
Adenocarcinoma	76	9869	(9869/0/0)
Prostate	102	6033	(6033/0/0)
Breast2	77	4869	(4869/0/0)
Leukemia	38	3052	(3052/0/0)
Colon	61	2000	(2000/0/0)
nki70 Breast Cancer	144	77	(72/1/4)
Glaucoma M	198	62	(62/0/0)
Wpbc	194(198)	33	(31/2/0)
Body	507	24	(24/0/0)
Biopsy	683(699)	9	(0/9/0)
SAheart	462	9	(5/3/1)
Diabetes	768	8	(8/0/0)
Appendicitis	106	7	(7/0/0)
Bupa	345	6	(1/5/0)
Dystrophy	194	5	(2/3/0)
Mammographic	830(961)	5	(0/5/0)
Transfusion	748	4	(2/2/0)
Hepatitis	80	19	(2/17/0)
Indian Liver Patients	583	10	(5/4/1)
Haberman	306	3	(0/3/0)
Phoneme	1000	5	(5/0/0)
Two Norms	1000	20	(20/0/0)
German Credit	1000	20	(0/7/13)
House voting	435	16	(0/0/16)
Bands	365	19	(13/6/0)
Sonar	208	60	(60/0/0)

In case of non-informative features in the data sets from Table 3, the ESkNN outperforms kNN based methods on most of the data sets. Comparing to random forest it gives low Brier scores on 10 data sets. These results indicate that the ESkNN is better than the kNN and kNN based methods and comparable to random forest.

ESkNN is evaluated for various values of k, the number of nearest neighbours and m, the number of models in the initial ensemble. Figure 2 reveals varied behaviour of ESkNN on different data sets for the choice of k and m. It is recommended to fine tune the value of k by cross validation, for example. Figure 2 (b) shows that a very small number of models are not reasonable and a very large number of models might be computationally expensive hence a moderate number of models is recommended.

4.2 Simulation Study

We evaluate the predictive performance of ESkNN by simulation study in addition to the benchmark data sets. We used two examples in our simulation study. The models proposed in our simulation study involve several variations to gain an understanding of the behaviour of the methods under different situations.

Simulation Model 1

In the first model, Model 1, binary class data is generated on 20 features. The features for class 1 are generated from $\mathcal{N}(\mathbf{2}, w\mathbf{\Psi})$, while those of class 2 are

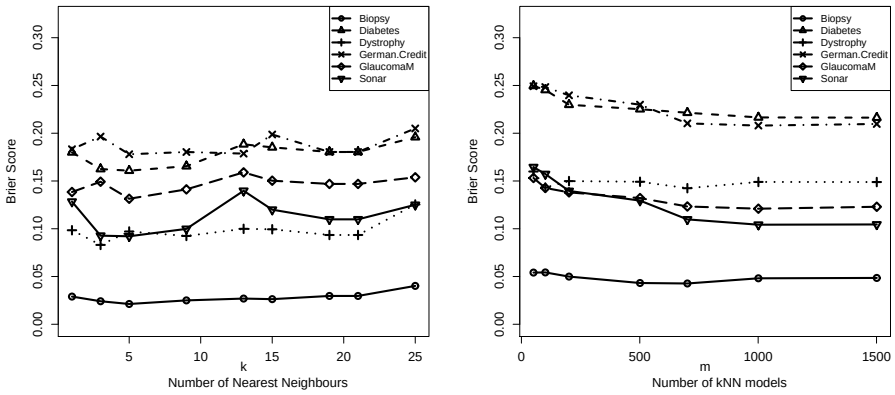
Table 2. Brier scores on the data sets on five methods.

Data Sets	*k*NN	B*k*NN	R*k*NN	ES*k*NN	RF
Haberman	0.199	0.197	0.181	**0.171**	0.199
Dystrophy	0.105	0.102	0.098	0.097	0.096
Mammographic	0.141	0.140	0.127	**0.115**	0.129
Transfusion	0.168	0.167	0.164	**0.160**	0.172
Bupa	0.221	0.215	0.217	0.215	**0.190**
Appendicitis	0.126	0.119	0.109	**0.105**	0.119
Diabetes	0.177	0.172	0.173	0.168	**0.156**
Biopsy	0.024	0.024	0.025	**0.021**	0.025
SAheart	0.209	0.207	0.203	0.200	**0.189**
Bands	0.235	0.231	0.207	0.208	**0.183**
German Credit	0.216	0.214	0.201	0.178	**0.159**
Body	0.019	0.019	0.036	**0.012**	0.031
Wpbc	0.182	0.182	0.176	0.172	**0.168**
Sonar	0.179	0.179	0.109	**0.092**	0.127
Glaucoma M	0.147	0.144	0.142	0.130	**0.089**
Indian liver	0.191	0.189	0.179	**0.163**	0.174
Phoneme	0.130	0.128	0.130	0.121	**0.105**
Two Norms	0.067	0.068	0.084	**0.029**	0.062
Hepatitis	0.310	0.259	0.209	0.221	**0.195**
House voting	0.065	0.065	0.065	0.044	**0.030**
Colon	0.145	0.144	0.139	0.138	**0.129**
Leukaemia	0.030	0.030	0.062	**0.027**	0.054
Breast2	0.243	0.241	0.233	0.230	**0.210**
Prostate	0.138	0.137	0.145	0.100	**0.084**
Adenocarcinoma	0.126	0.119	0.119	**0.114**	0.125

Table 3. Brier scores of the methods with added non-informative features to the data.

Data Sets	*k*NN	B*k*NN	R*k*NN	ES*k*NN	RF
Haberman	0.204	0.202	0.196	**0.191**	0.196
Dystrophy	0.158	0.172	0.220	0.149	**0.118**
Mammographic	0.153	0.160	0.231	0.139	**0.123**
Transfusion	0.187	0.186	0.180	**0.160**	0.166
Bupa	0.229	0.228	0.243	**0.222**	0.230
Appendicitis	0.143	0.142	0.145	0.139	**0.132**
Diabetes	0.240	0.236	0.225	0.216	**0.173**
Biopsy	0.053	0.052	0.067	0.048	**0.029**
SAheart	0.252	0.247	0.228	0.225	**0.218**
Bands	0.237	0.235	0.222	**0.213**	0.221
German Credit	0.218	0.216	0.210	0.208	**0.182**
Body	0.082	0.082	0.107	0.078	**0.065**
Wpbc	0.196	0.190	0.180	**0.179**	0.181
Sonar	0.164	0.139	0.201	**0.104**	0.193
Glaucoma M	0.157	0.156	0.212	**0.121**	0.135
Indian liver	0.198	0.199	0.201	**0.183**	0.189
Phoneme	0.174	0.170	0.236	**0.154**	0.163
Two Norms	0.126	0.084	0.203	**0.082**	0.124
Hepatitis	0.239	0.230	0.239	**0.223**	0.234
House Voting	0.135	0.134	0.212	0.127	**0.103**

generated from $\mathcal{N}(\mathbf{1}, \mathbf{1})$. The values considered for w in class 1 are 3, 5, 10, 15 and 20. The predictive performance of the algorithms are investigated by adding 50, 100, 200 and 500 non-informative features, generated from normal



(*a*) (*b*)

Fig. 2. Performance of ESkNN in presence of non-informative features in the data for; (a): different values of k, (b): different values of m

distribution, to the data.

$$\mathbf{\Psi} = \begin{pmatrix} \sigma_{1,1} & \varrho_{1,2} & , ..., & \varrho_{1,d} \\ \varrho_{2,1} & \sigma_{2,2} & , ..., & \varrho_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ \varrho_{d,1} & \varrho_{d,2} & , ..., & \sigma_{d,d} \end{pmatrix} \tag{4}$$

where ϱ_{ij} are the covariances between the features defined as:

$$\varrho_{ij} = (1/2)^{|i-j|}, i, j = 1, ..., d, \tag{5}$$

and $\sigma_{ij} = 1$ for $i = j$. The variables within class 1 are correlated among each other and are exhibiting negligible/no correlation with the features of class 2.

Simulation Model 2

The second simulation model developed here is a four-dimensional model, derived from the model proposed by Mease et al. (2007). The feature vector $\mathbf{x}$ is a random vector uniformly distributed over $[0, 100]$. The class is determined by the distance r, the distance of the feature vector $\mathbf{x}$ from the central point. The class probabilities given the features are:

$$p(y = 1 \mid \mathbf{x}) = \begin{cases} 1 & r < 110, \\ \frac{150-r}{140} & 110 \leq r \leq 140, \\ 0 & \textit{otherwise}. \end{cases}$$

The binary response variable y is generated from the above distribution using a binomial random number generator. We extended the dimensions of the

data by adding 50,100, 200 and 500 randomly generated non-informative feature.

4.3 Simulation Results and Discussion

The results from Table 4 reveal that ES*k*NN consistently outperform the other methods. In case of different values of w to the data in Model 1, as shown in Table 4, random forest outperforms all the other methods. However, in *k*NN based methods the ES*k*NN consistently gives higher accuracy than *k*NN, B*k*NN and R*k*NN.

The Brier scores from Model 2 given in Table 5, show that ES*k*NN consistently outperforms *k*NN, B*k*NN, R*k*NN and RF for the data with original 4 features and added 50, 100, 200 and 500 features.

Table 4. Brier score of the five methods with added non-informative features to the data set from Model 1 and different values of w on 70 features (20+ 50 non-informative) shown in first column. The best result is highlighted in bold.

Features	***k*NN**	**B*k*NN**	**R*k*NN**	**ES*k*NN**	**RF**
20	0.042	0.041	0.087	**0.039**	0.071
20+50	0.066	0.079	0.086	**0.060**	0.081
20+100	0.081	0.076	0.095	**0.061**	0.086
20+200	0.103	0.094	0.095	**0.062**	0.092
20+500	0.137	0.130	0.088	**0.061**	0.113
w	***k*NN**	**B*k*NN**	**R*k*NN**	**ES*k*NN**	**RF**
3	0.198	0.151	0.155	0.102	**0.081**
5	0.221	0.191	0.136	0.101	**0.062**
10	0.222	0.186	0.099	0.081	**0.038**
15	0.251	0.172	0.089	0.057	**0.028**
20	0.256	0.159	0.062	0.043	**0.022**

Table 5. Brier score of the methods on the data from Model 2 with the added non-informative features. Results of the best performing method is highlighted in bold.

Features	***k*NN**	**B*k*NN**	**R*k*NN**	**ES*k*NN**	**RF**
4	0.101	0.101	0.145	**0.090**	0.112
4+50	0.158	0.157	0.185	**0.146**	0.176
4+100	0.165	0.164	0.190	**0.152**	0.186
4+200	0.179	0.178	0.196	**0.162**	0.177
4+500	0.188	0.182	0.209	**0.151**	0.180

5 Conclusion

We proposed an ensemble of subset of *k*NN models, ES*k*NN, for class membership probability estimation. The ES*k*NN improves the predictive performance of *k*NN based methods. The ES*k*NN reveals better predictive performance than the *k*NN, bagged *k*NN and random *k*NN in most of the cases (both in bench marking and simulation) and gives comparable results to random forest. The performance of ES*k*NN is also evaluated in order to deal with the

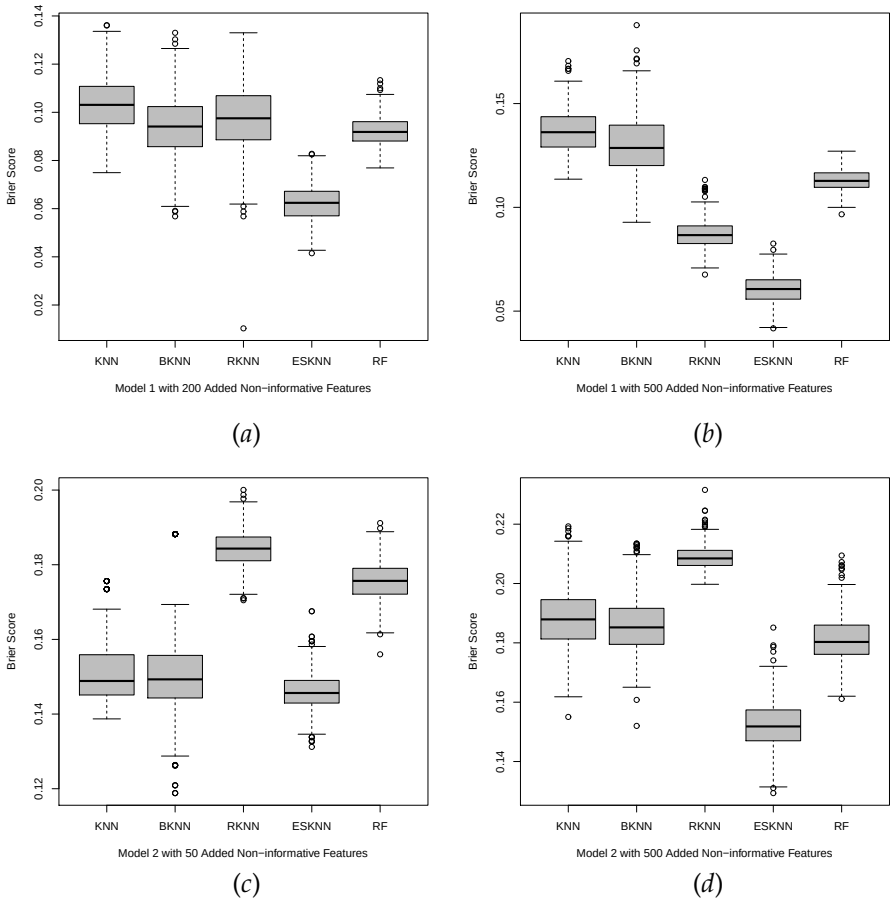


Fig. 3. Brier score, of simulated data from Model 1 and Model 2 for the five classifiers *k*NN, B*k*NN, R*k*NN, ES*k*NN and RF with added non-informative features to the data.

issue of non-informative features in the data. The results demonstrate that the ES*k*NN provides better estimates of class membership probability than the other methods considered in the presence of non-informative features in the data. Besides performance improvement, the ES*k*NN as using *k*NN classifier is simple in implementation and interpretation. The ES*k*NN is implemented in an R-package, **ES*k*NN**.

References

BAY, S. (1998). Combining Nearest Neighbor Classifiers through Multiple Feature Subsets. In Proceedings of the Fifteenth International Conference on Machine Learning, 3, 37–45.

BREIMAN, L. (1996): Bagging Predictors. Machine Learning, 24(2), 123–140.

BRIER, G. W. (1950): Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78, 1-3.

COVER, T. and HART, P. (1967): Nearest Nieghbor Pattern Classification. IEEE Transaction on Information Theory, 13, 21–27

GNEITING, T. and RAFTERY, A. E. (2007): Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102, 359-378.

HOTHORN, T. and LAUSEN, B. (2003): Double-Bagging: Combining Classifiers by Bootstrap Aggregation. Pattern Recognition, 36 (9), 1303–1309.

KHAN, Z., PERPEROGLOU, A., GUL, A., MAHMOUD, O., ADLER, W., MiFTAHUDDIN, M. and LAUSEN, B. (2015): An Ensemble of Optimal Trees for Class membership Probability Estimation. In Proceedings of European Conference on Data Analysis.

KRUPPA, J., ZIEGLER, A. and KONIG, I. R. (2012): Risk Estimation and Risk Prediction Using Machine-Learning Methods. Human Genetics, 131, 1639-1654.

KRUPPA, J., LIU, Y., BIAU, G., KOHLER, M., KONIG, I. R., MALLEY, J. d. and ZIEGLER, A. (2014a): Probability Estimation with Machine Learning Methods for Dichotomous and Multicategory Outcome: Theory. Biometrical Journal, 56, 534-563.

KRUPPA, J., LIU, Y., DIENER, H. C., WEIMAR, C., KONIG, I. R. and ZIEGLER, A. (2014b): Probability Estimation with Machine Learning Methods for Dichotomous and Multicategory Outcome: Applications. Biometrical Journal, 56, 564-583.

KUNCHEVA, L. I.(2004): Combining Pattern Classifiers. Methods and Algorithms. *John Wiley and Sons.*

LEE, B. K., LESSLER, J. and STUART, E. A. (2010): Improving Propensity Score Weighting using Machine Learning. Statistics in Medicine, 29, 337-346.

LI, S., HARNER, E. J. and ADJEROH, D. (2011): Random knn Feature Selection a Fast and Stable Alternative to Random Forests. BMC bioinformatics, 12(1), 450.

MAHMOUD, O., HARRISON, A., PERPEROGLOU, A., GUL, A., KHAN, Z., METODIEV, M. V. and LAUSEN, B. (2014a): A Feature Selection Method for Classification within Functional Genomics Experiments based on the Proportional Overlapping Score. BMC Bioinformatics, 15, 274.

MAHMOUD, O., HARRISON, A., PERPEROGLOU, A., GUL, A., KHAN, Z. and LAUSEN, B. (2014b): propOverlap: Feature (gene) selection based on the Proportional Overlapping Scores. R package version 1.0, `http://CRAN.R-project.org/package=propOverlap`.

MALLEY, J., KRUPPA, J., DASGUPTA, A., MALLEY, K. and ZIEGLER, A. (2012): Probability Machines: Consistent Probability Estimation using Nonparametric Learning Machines. Methods of Information in Medicine, 51, 74–81.

MEASE, D., WYNER, A. J. and Buja, A. (2007): Boosted Classification Trees and Class Probability/Quantile Estimation. The Journal of Machine Learning Research, 8, 409–439.

MELVILLE, P., SHAH, N., MIHALKOVA, L. and MOONEY, R. (2004): Experiments on ensembles with missing and noisy data. Multiple Classifier Systems, 293–302.

*NETTLETON, D. F., ORRIOLS-PUIG, A. and FORNELLS, A. (2010): A Study of the Effect of Different Types of Noise on the Precision of Supervised Learning Techniques.*Artificial intelligence review, 33(4), 275–306.

SAMWORTH, R. J. (2012): Optimal Weighted Nearest Neighbour Classifiers. The Annals of Statistics, 40(5), 2733–2763.