Choosing Parameters for Random Subspace Ensembles for fMRI Classification

Ludmila I. Kuncheva and Catrin O. Plumpton

School of Computer Science, University of Bangor, UK
{l.i.kuncheva,c.o.plumpton}@bangor.ac.uk

Abstract. Functional magnetic resonance imaging (fMRI) is a non-invasive and powerful method for analysis of the operational mechanisms of the brain. fMRI classification poses a severe challenge because of the extremely large feature-to-instance ratio. Random Subspace ensembles (RS) have been found to work well for such data. To enable a theoretical analysis of RS ensembles, we assume that only a small (known) proportion of the features are important to the classification, and the remaining features are noise. Three properties of RS ensembles are defined: usability, coverage and feature-set diversity. Their expected values are derived for a range of RS ensemble sizes (L) and cardinalities of the sampled feature subsets (M). Our hypothesis that larger values of the three properties are beneficial for RS ensembles was supported by a simulation study and an experiment with a real fMRI data set. The analyses suggested that RS ensembles benefit from medium M and relatively small L.

1 Introduction

Functional magnetic resonance imaging (fMRI) measures blood oxygenation level-dependent (BOLD) signal in a quest to discover how mental states are mapped onto patterns of neural activity. Advanced as they are, pattern recognition and machine learning are yet to contribute powerful bespoke techniques to fMRI data analysis [1, 2]. The formidable challenges come from: (i) the extremely large feature-to-instance ratio, in the order of 5000:1; (ii) the spatial relationship between the features (voxels in the 3-D image of the brain); (iii) the low contrast-to-noise ratio; and (iv) the great redundancy in the feature set. Preferences tend to be for linear classifiers because they are simple, fast, reasonably accurate and interpretable. The favourite, however, has been the support vector machine classifier (SVM) [3–7]. A recent comparison of classification methods for an fMRI data set placed the Random Subspace Ensemble (RS) with SVM base classifiers as the most accurate classification method across a variety of voxel pre-selection methods [8]. To construct a random subspace ensemble with L classifiers, L samples of size M are drawn without replacement from a uniform distribution over the set of voxels. A classifier is trained on each feature subset using either the whole training set or a bootstrap sample thereof [9].

N. El Gayar, J. Kittler, and F. Roli (Eds.): MCS 2010, LNCS 5997, pp. 54–63, 2010.

Random Subspace ensembles have been considered for problems with large dimensionality and excessive feature-to-instance ratio [10], e.g., problems arising from microarray data analysis [11] and face recognition [12]. The overwhelming computational demand in applying RS to the raw fMRI data led to the idea of pre-selection of voxels. Univariate statistical methods have been employed for that [13, 14]. The importance of a voxel is measured by the p-value of a t-test (or ANOVA for multiple classes) for equivalence of the class means. The initial set of voxels is subsequently reduced to a subset of 1000 or 2000 voxels, and RS ensembles are created on that subset. Admittedly, univariate approaches may destroy important relationships between features. Such features would not be indicative individually but may form a highly indicative group. In balance, pre-selection eliminates the vast majority of irrelevant voxels which justifies some (hypothetical) loss of discriminative information.

Relevance and redundancy are two different aspects in large-scale feature selection [15], and both are present in fMRI data. As the data is a "snapshot" of the whole brain, the vast majority of the voxels are irrelevant for each particular task. The relevant voxels, on the other hand, are likely to be spatially grouped into clusters exhibiting large correlations (redundancy). Most voxel selection methods do not guard against redundancy because the position, size and shape of the clusters of relevant voxels is of interest to the investigator. Thus we examine the effect of the number of relevant voxels on the parameter choices of RS ensembles. An advantage of RS ensembles compared to many other ensemble methods and single classifiers is that they need only two parameters, L, the ensemble size and M, the size of the feature sample. Given the specifics of fMRI data, this paper offers a theoretical perspective on choosing values of these parameters. Section 2 introduces the theoretical framework. Simulation experiments are reported in Section 3, and discussed in Section 4.

2 Random Subspace Ensembles

Let $X = \{x_1, \ldots, x_n\}$ be the set of n features (voxels). L samples, each of size M, are drawn without replacement from a uniform distribution over X and a classifier is trained on each sample. The ensemble decision is made by majority vote among the L classifiers.

In many fMRI studies, the relevant information is typically a sparse irregular pattern of responsive voxels in the 3-D image of the brain. It is likely that a small number of voxels contain most of the discriminative information. We assume that there are Q "important" voxels, set $\mathcal{I} = \{q_1, \ldots, q_Q\}$, $\mathcal{I} \subset X$, where $|\mathcal{I}| = Q << n$, and the remaining $n - Q$ voxels are random noise. We also assume that the cardinality of the subspaces, M, is much smaller than n. The question is whether we can recommend L and M for a given n and Q. We base this study on the following postulate [16–18].

Postulate. Accurate and diverse individual classifiers are a prerequisite for better ensembles. ■

The subset of features, on which the individual classifiers are built, can serve as an indication of the expected accuracy and diversity of these classifiers. If a classifier uses only 'noise' features, its accuracy will be no better than random chance. Also, classifiers that use the same 'important' features will be similar or identical, therefore redundant in the ensemble. Finally, we would like the whole of $\mathcal{I}$ to be covered, so that important information is not lost. In other words, we would like each $q \in \mathcal{I}$ to be selected at least once in the L samples of M features.

2.1 Usability

Definition 1. We call a classifier *usable* if its feature subset contains at least one 'important' voxel $q \in \mathcal{I}$. ■

To calculate the probability of drawing a feature subset of a usable classifier, take Y to be the number of 'important' features in a subset of size M, drawn without replacement from X. Y is a random variable with hypergeometric distribution with probability mass function

$$P(Y=i) = \frac{\binom{Q}{i}\binom{n-Q}{M-i}}{\binom{n}{M}}, \quad i = 0, 1, \ldots, Q.$$

The probability of drawing a usable classifier is

$$P(\text{usable classifier}) = 1 - P(Y=0) = 1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}$$

Definition 2. The *degree of usability of the ensemble*, U, is defined as the proportion of usable classifiers out of L. ■

Since the subsets are sampled independently, the probability of having a completely usable ensemble is

$$P(U=1) = P(\text{usable classifier})^L = \left(1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}\right)^L. \tag{1}$$

The ratio of the two binomial coefficients can be simplified for computational purposes to give

$$P(U=1) = \left(1 - \prod_{i=0}^{M-1}\left(1 - \frac{Q}{n-i}\right)\right)^L. \tag{2}$$

Since we assumed $M << n$, the equation can be simplified further to

$$P(U=1) \approx \left(1 - \left(1 - \frac{Q}{n}\right)^M\right)^L. \tag{3}$$

This approximation is equivalent to replacing the hypergeometric distribution with a binomial distribution. Given the size of n for fMRI data, we can say that sampling *with* replacement is approximately equivalent to sampling *without* replacement. Y can therefore be approximated with a binomial distribution with parameters M and $p = \frac{Q}{n}$. The probability of a usable classifier in this case would be $1 - \left(1 - \frac{Q}{n}\right)^M$.

To calculate the expected value of the degree of usability of the ensemble, $E[U]$, let Z be a random variable expressing the number of usable classifiers in the ensemble. Z has a hypergeometric distribution with the following parameters. The *total* is the number of all possible samples (without replacement) of size M from X, i.e., $\binom{n}{M}$. The number of *usable* classifiers is calculated by taking the number of non-usable classifiers, $\binom{n-Q}{M}$, from the total. The number of *selected* classifiers at a time is L. The expected value of Z is $\frac{\text{Selected} \times \text{Usable}}{\text{Total}}$, therefore the expected usability of the ensemble is $E[U] = \frac{1}{L}E[Z]$

$$E[U] = \frac{1}{L} \times L \times \left(1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}\right) = 1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}. \tag{4}$$

The expected usability of the ensemble is equivalent to the probability of selecting a usable classifier, and does not depend on the ensemble size L. Our hypothesis is that the higher the degree of usability, the more accurate the ensemble.

2.2 Coverage

Definition 3. The *degree of coverage of the ensemble*, C, is the proportion of the features $q \in \mathcal{I}$ (out of Q) selected for one of more of the base classifiers. ■

For calculating coverage, we can again use the binomial approximation to the hypergeometric distribution. This implies that the feature subsets are sampled independently from the X. For a given important feature $q \in \mathcal{I}$ the probability of selecting that feature in a sample of size M is $\frac{M}{n}$. The probability of not selecting q in a sample of size M is therefore $1 - \frac{M}{n}$. The probability of not selecting q in at least one of L samples of size M is $P(\bar{q}) = \left(1 - \frac{M}{n}\right)^L$. The probability of q being selected in at least one of the L samples is $1 - P(\bar{q})$. The probability of all features being covered is

$$P(\text{Complete coverage}) = P(C = 1) = \left(1 - \left(1 - \frac{M}{n}\right)^L\right)^Q. \tag{5}$$

Denote by Z the number of covered features out of Q. Z has binomial distribution with parameters Q and $p = 1 - \left(1 - \frac{M}{n}\right)^L$. The expected coverage is

$$E[C] = \frac{1}{Q}\left(1 - \left(1 - \frac{M}{n}\right)^L\right) Q = 1 - \left(1 - \frac{M}{n}\right)^L. \tag{6}$$

The expected coverage depends on the ensemble size L and the subset size M but not on Q. The hypothesis here is that the higher the degree of coverage, the more accurate the ensemble.

2.3 Feature Set Diversity

Note that for fixed n and Q, $E[U]$ is monotonically increasing on M, and $E[C]$ increases with both L and M. This suggests that larger ensembles with larger feature sample size M should be preferred. In the extreme case where $M = n$, the ensemble will contain identical copies of the base classifier trained on all features. This defeats the point of having an ensemble altogether. Besides, with the extremely large feature-to-instances ratio, the individual classifier may easily overfit the data. Therefore we introduce a third property.

Definition 4. Denote by $S_1, S_2, \ldots, S_L$ the L feature subsets sampled from X. Consider $S_1, S_2 \subset X$ such that $|S_1| = |S_2| = M$. Denote by $I_1 \subseteq \mathcal{I}$ and $I_2 \subseteq \mathcal{I}$ the respective subsets of 'important' features in S_1 and S_2 respectively. We define *Feature Set Diversity* (D) as

$$D(S_1, S_2) = |I_1 \cup I_2| - |I_1 \cap I_2|. \qquad \blacksquare$$

Two classifiers are *non-identical* if their feature subsets differ by at least one 'important' voxel. Each feature $q \in \mathcal{I}$ may or may not contribute to D. A value of 1 will be added if q is in either set but not in both. Then the expected diversity for any pair of subsets S_1 and S_2 is

$$E[D] = \sum_{i=1}^{Q} P(q_i \in I_1)P(q_i \notin I_2) + P(q_i \notin I_1)P(q_i \in I_2).$$

Since all features in $\mathcal{I}$ have equal chance of being selected in a subset of size M, and the subsets are drawn independently,

$$E[D] = 2Q\frac{M}{n}\left(1 - \frac{M}{n}\right). \tag{7}$$

This calculation disregards non-usable classifiers. So an ensemble can be diverse even if it contains non-usable classifiers for which $I_1 = I_2 = \emptyset$.

Figure 1 shows the theoretical and simulated curves for $E[U]$ (4), $E[C]$ (6) and $E[D]$ (7) for $n = 1000$, $Q = 100$ and $L = 10$. Changing the value of L to 50 and 100, and Q to 10 and 50 did not lead to large differences in the shapes and positions of the curves. The results suggest that values of M close to $\frac{n}{2}$ are optimal as all three criteria reach their maxima, also observed across different ensemble sizes.

3 A Simulation Experiment

The important question here is to what extent the three characteristics are related to the classification accuracy of the RS ensemble.

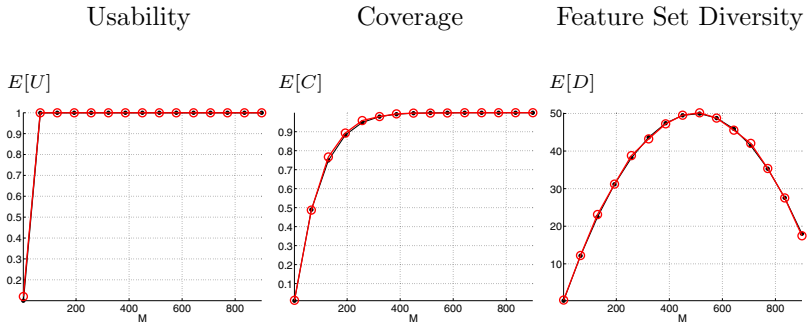

Fig. 1. Theoretical and simulation curves (coinciding) for the expected values of U, C and D for $n = 1000$, $Q = 100$ and $L = 10$. The empirical curve is calculated as an average of 10 ensembles with randomly sampled $L = 10$ sets of M features.

3.1 Data

We decided to use simulated data that exhibit properties similar to real data while keeping control on the parameters n and Q. We used an fMRI data set collected at the School of Psychology, University of Bangor. The data consisted of the single-subject BOLD responses to 3 types of stimuli: faces, places and objects. Each presentation of a stimulus defined a point in the data set. The total number of voxels (features) was 106720 and the number of objects was 36, 12 in each class. The classification task was to predict which type of stimuli the subject is looking at, judging by the fMRI response.

For a 2-class problem, Contrast-to-Noise-Ratio (CNR) is defined using the means and the standard deviations for the classes, separately for each voxel. For voxel v, CNR is $\frac{\mu_1(v)-\mu_2(v)}{\frac{1}{2}(\sigma_1(v)+\sigma_2(v))}$, where $\mu_i(v)$ is the mean and $\sigma_i(v)$ is the standard deviation of v for class i. The higher the CNR, the more separable the two classes are using only voxel v. We took only classes 1 and 2 (faces and places) and calculated CNR for each voxel. The voxels were then sorted by their CNR, in descending order. The means and the covariance matrices for the two classes of the top Q voxels were stored and subsequently used to simulate the first (important) Q features in the data. We simulated multivariate Gaussian distributions for each class, using the Statistics toolbox of Matlab. The remaining $n - Q$ features were simulated as independent random noise with mean zero and standard deviation equal to the mean CNR for the Q important features. Running a separate simulation study even in addition to the experiments with the real data was necessary in order to have *control over* Q in a surrogate pseudo-real environment.

3.2 Experimental Protocol

The parameters were varied in the following ranges: the total number of features, n, took values 200, 500 and 1000, and the number of 'important' features, Q,

was chosen accordingly to model ratios $\frac{Q}{n}$ of 0.02, 0.05, 0.1, 0.25, 0.5 and 1. The feature set cardinality, M, took 20 equally spaced values from 1 to n, and the ensemble size L took values at regular intervals from 1 to 200.

For each combination (M, L, Q, n), we generated 10 data sets with 20 training examples (10 per class) and 200 testing examples (100 per class). The small size of the training data was chosen to mirror that of real data sets. SVM was used as the base classifier. For each (M, L, Q, n) we calculated the RS ensemble error, and also estimated the observed degree of usability U, the degree of coverage C, and the feature set diversity D of the ensemble.

3.3 Results

Figure 2 gives an example of the type of surfaces over the (L, M) grid, obtained through the simulations. Each point in the space is calculated as the average across 10 simulations with data drawn independently from the chosen 'realistic'

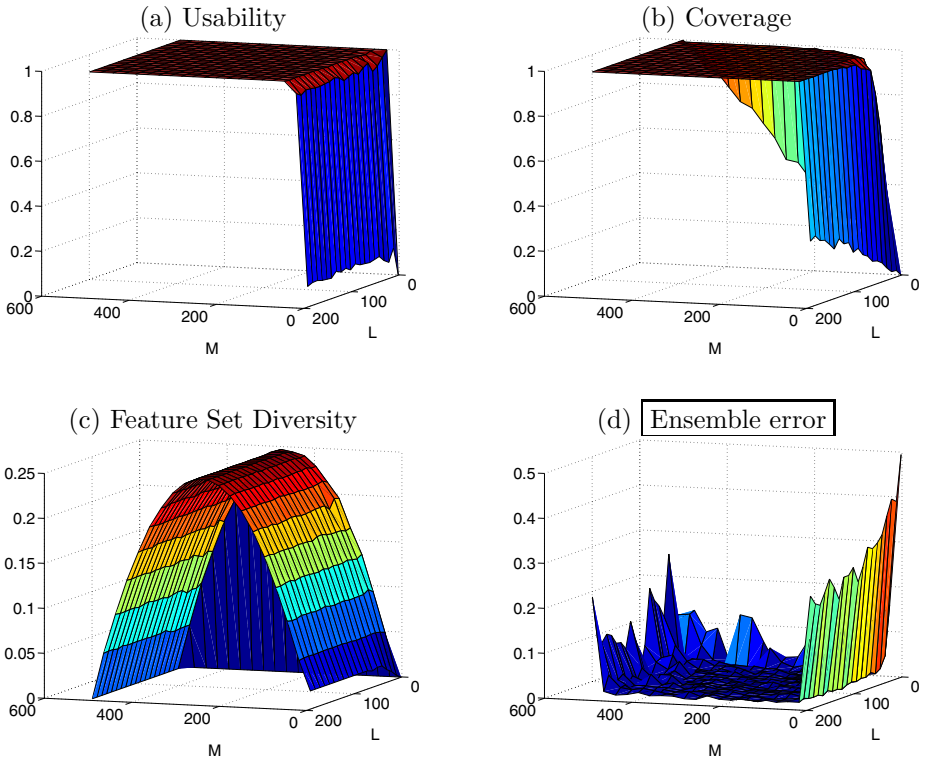

Fig. 2. The three RS characteristics and the ensemble error as functions of the ensemble size L and the feature set size M. Each of the 2 classes in the data set was sampled from a Gaussian distributions with $Q = 50$ relevant and $n - Q = 450$ noise features.

Table 1. Summary of the simulation results. $\bar{E}$ is the average RS ensemble error across the (L, M) grid. $\bar{E}^*$ is the value of the ensemble error for the recommended parameter values, $M = \frac{n}{2}$ and $L = \frac{n}{10}$.

			Correlation with E		
$\frac{Q}{n}$ ratio	$\bar{E}$	$\bar{E}^*$	Usability	Coverage	Diversity
0.02	4.09	1.25	−0.940	−0.863	−0.410
0.05	3.81	1.75	−0.972	−0.903	−0.438
0.10	3.83	1.40	−0.855	−0.802	−0.581
0.25	3.19	0.90	−0.593	−0.640	−0.608
0.50	3.65	2.15	−0.123	−0.233	−0.608
1.00	6.88	1.05	N/A	0.014	−0.448

distribution. The surfaces in the plots were obtained with $n = 500$ and $Q = 50$. They have typical shapes observed for the six $\frac{Q}{n}$ ratios. The surfaces confirm visually the hypothesis that larger usability, coverage and feature set diversity lead to better ensembles (lower error).

As expected, the degree of usability (subplot (a)) does not depend on L, and quickly raises to 1 with M. The 'tent'-shaped surface of D in subplot (c) also depends on M but not on L. Largest values of D are achieved for $M \approx \frac{n}{2}$. Like U, the degree of coverage, C, maintains its maximum value of 1 for the largest part of the (L, M) grid. The figure also suggests that small to medium values of the ensemble size L are sufficient. Hence, as a rule of thumb, we recommend $M = \frac{n}{2}$ and $L = \frac{n}{10}$ for fMRI type of data.

Table 1 shows a summary of the simulation results. We show the $\frac{Q}{n}$ ratio, the average error rate of the RS ensemble over the whole (L, M) grid, $\bar{E}$, as well as the error using the *recommended* values, $\bar{E}^*$. For all $\frac{Q}{n}$ ratios, $\bar{E} > \bar{E}^*$. We also give the correlation coefficients between the RS ensemble error E, on the one hand, and U, C, and D, on the other hand. Even though calculated on an artificial grid, these coefficients support the hypothesis that large values of usability, coverage and feature-set diversity are beneficial for the ensemble.

3.4 Experiment with the Real fMRI Data

The RS ensemble with SVM base classifiers was run on classes 1 and 2 (faces and places) of the real fMRI data set. First, $n = 1000$ voxels were pre-selected by the SVM method [14]. An SVM classifier was trained on all voxels, the voxels were sorted by descending absolute value of the SVM weights, and the top 1000 voxels were retained. Three-fold cross-validation was applied to test the RS ensemble for a 10×10 grid of values for M and L. M was varied from 1 to n at equal intervals, and L was varied from 1 to $n/5$. Figure 3 plots the surface of the ensemble error over the (L, M) grid. The recommended values of $M = 500$

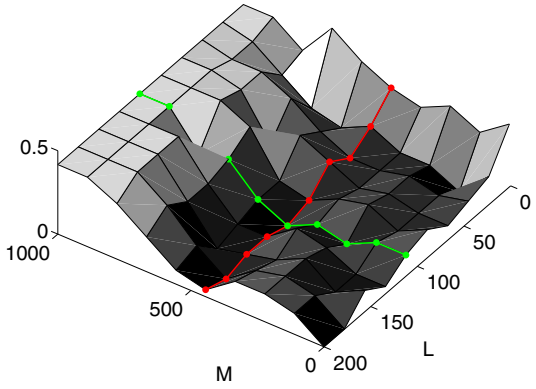

Fig. 3. RS error on the real fMRI data set as a function of the ensemble size L and the feature size M. The recommended values for L and M are overlayed on the surface.

and $L = 100$ are marked as lines on the 3-D plot. The lines intersect near the minimum of the error surface, which confirms empirically the recommendation for L and M. While the average error over the whole grid was 0.2138, the error at $M = 500$ and $L = 100$ was 0.0521.

4 Conclusions

We examine the Random Subspace ensemble (RS) for fMRI type of data where the feature-to-instance ratio is in the order of 50000:1, and the number of truly relevant features (voxels in the 3-D image of the brain) is much smaller than the total number of features. Following previous fMRI studies we consider n pre-selected voxels, where n is in the region of 1000. Assuming that there are Q 'important' features among the pre-selected n features, three characteristics of the RS ensemble are defined: usability U, coverage C and feature-set diversity D. Expected values of these characteristics are derived theoretically as functions of n, Q, L and M. Our hypothesis was that higher values of U, C and D are beneficial for the RS ensemble. A simulation study was carried out, with two heteroscedastic Gaussian classes whose covariance matrices were estimated from a real fMRI data set and augmented with Gaussian noise. The results support the research hypothesis. As a rule of thumb, we propose to use feature set size $M = \frac{n}{2}$ and ensemble size $L = \frac{n}{10}$. These values were found to work well for the real fMRI data.

Acknowledgements

We are grateful to David Linden and Stephen Johnston, School of Psychology, Bangor University, UK, for providing the fMRI data.

References

1. Norman, K.A., Polyn, A.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10, 424–430 (2006)
2. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45(1, suppl. 1), 199–209 (2009)
3. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fMRI): detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19(2), 261–270 (2003)
4. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fMRI data. NeuroImage 26(2), 317–329 (2005)
5. Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M.: Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage 28(4), 980–995 (2005)
6. Wang, Z., Childress, A.R., Wang, J., Detre, J.A.: Support vector machine learning-based fMRI data group analysis. NeuroImage 36(4), 1139–1151 (2007)
7. Ku, S.-p., Gretton, A., Macke, J., Logothetis, N.K.: Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magnetic Resonance Imaging 26(7), 1007–1014 (2008)
8. Kuncheva, L.I., Rodríguez, J.J.: Classifier ensembles for fMRI data analysis: An experiment. Magnetic Resonance Imaging (to appear)
9. Ho, T.K.: The random space method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
10. Skurichina, M., Duin, R.P.W.: Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications 5, 121–135 (2002)
11. Lai, C., Reinders, M.J., Wessels, L.: Random subspace method for multivariate feature selection. Pattern Recognition Letters 27(10), 1067–1076 (2006)
12. Zhu, Y., Liu, J., Chen, S.: Semi-random subspace method for face recognition. Image and Vision Computing 27(9), 1358–1370 (2009)
13. Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (eds.): Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, London (2007)
14. De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E.: Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43(1), 44–58 (2008)
15. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
16. Kuncheva, L.I.: Combining Pattern Classifiers. Methods and Algorithms. John Wiley and Sons, New York (2004)
17. Brown, G.: Ensemble learning. In: Sammut, C., Webb, G. (eds.) Encyclopedia of Machine Learning. Springer, Heidelberg (2009)
18. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Information Fusion 6(1), 5–20 (2005)