Cancer Classification Using Ensemble of Error Correcting Output Codes

Zhihao Zeng, Kun-Hong Liu, and Zheyuan Wang

Xiamen University, School of Software
{zengzhihao5star,wang1987}@gmail.com, lkhqz@xmu.edu.cn

Abstract. We address the microarray dataset based cancer classification problem using a newly proposed ensemble of Error Correcting Output Codes (***E-ECOC***) method. To the best of our knowledge, it is the first time that ECOC based ensemble has been applied to the microarray dataset classification. Different feature subsets are generated from datasets as inputs for some problem-dependent ECOC coding methods, so as to produce diverse ECOC coding matrixes. Then, the mutual difference degree among the coding matrixes is calculated as an indicator to select coding matrixes with maximum difference. Local difference maximum selection(***L-DMS***) and global difference maximum selection(***G-DMS***) are the strategies for picking coding matrixes based on same or different ECOC algorithms. In the experiments, it can be found that E-ECOC algorithm outperforms the individual ECOC and effectively solves the microarray classification problem.

Keywords: ECOC, ensemble learning, Cancer classification, feature selection.

1 Introduction

In the field of machine learning and pattern recognition, the goal of a classification problem is looking for a map function: f: S – K, in which S is a set of attributes that describes series of properties of the samples, and K is the corresponding labels that belong to each sample. Function f maps each sample belonging to S into a unique class label k. Consider a binary problem, there has been widespread application of mature machine learning algorithm for estimating the function f. However, for multi-class problems, with the increasing categories, a single learner is usually hardly competent to produce accurate outputs. |And there are many classifier that can only deal with binary class problem. An alteration for solving the multi-class problem is the divide and conquer method, which means, the original classification problem is decomposed into multiple binary classification problems. By solving each binary classification independently, we can solve a multi-class classification task with some integration strategies, such as voting. Under the guidance of this idea, there are three basic solutions: flat strategy, hierarchical strategy, Error-Correcting Output Codes (ECOC). In flat strategy, a fixed decomposition method is used, such as One vs. One or One vs. All, and the final label is decided directly by voting. On the other hand,

D.-S. Huang et al. (Eds.): ICIC 2014, LNBI 8590, pp. 15–24, 2014.

hierarchical strategy build a binary tree based on the relationship among categories for the multi-class problem, and each branch node represents a binary classifier and the leaf node represents a final class. ECOC algorithm framework[1] consists of two key steps: in encoding phase, the original multi-classification problem is decomposed into multiple binary classification problem, which is represented by an M*N encoding matrix δ. In a coding matrix, each row represents a unique class, and each column illustrates specifically the decomposition method from a multi-class problem into a set of binary problem. In decoding phase, by comparing the distance between outputs of the multiple binary classifier and each code word in the coding matrix, the label with the minimum distance is selected as the final label for a unknown sample[2]. In a sense, ECOC algorithm framework can be considered as a more general solution than flat and hierarchical strategies. In the coding phase, the methods of decomposing multi-class contain all of the possible ways of division from the former two strategies. In addition, Dietterich and Kong[3] proved that ECOC algorithm framework can reduce bias and variance errors produced by the binary classification algorithms. It's worth noting that the number of the binary classifiers has been reduced to [$10\log_2$ N, 15 $\log_2$ N][4]. The coding matrix is not difficult to construct even when N is large enough. However, it is very difficult to filter the optimum coding matrix.

In the past few years, the ECOC algorithm framework were studied by researchers from different perspectives. Algorithms to construct suitable and effective coding matrix, and the decoding strategies have been extensively studied. Moreover, Masulli and Valentini[5] analyzed the different factors that affect the effectiveness of ECOC algorithm, and the correlation between the coding matrix and the binary classifier. Effectiveness of ECOC depends on code word correlation, structure and accuracy of dichotomizers, and the complexity of the multiclass learning problem. It is noticeable that the predefined coding matrix, like one vs. one, one vs. all, and the random-based coding matrix, are not suitable for the problems. The reason is that all those algorithms neglect the distribution characteristics of the data itself. Therefore, researchers take the distribution features of the data into consideration when constructing the encoding matrix and proposed many data dependent encoding algorithms to decompose the original multi-class problem into dichotomizer. DECOC[6] method builds N-1 binary classifier. Moreover, Crammer and Singer[7] proved that searching for the optimal coding matrix which are associated to the problem domain is a NP-complete problems. Recent research works use Genetic Algorithms in the coding phase to obtain higher accuracy of coding matrix along with reducing the number of dichotomizer. Bautista et al.[8] focused on optimizing ECOC coding matrix based on the standard genetic algorithm (GA), which is known as Minimal ECOC. The final result was that the number of binary classifiers is reduced to [$\log_2$ N] and at the same time, the degree of differentiation among classes are guaranteed. Garcia-Pedrajas and Fyfe[9] used the CHC based genetic algorithm to optimize the Sparse Random ECOC Matrix. In their work, the length of coding matrix is limited within [30, 50], and is independent of both the distribution of data sets and the number of classes. It is obvious that the techniques involved are simple and direct. Lorena and Carvalho [10]combined GA with the Sparse Random coding matrix too, and limited the length of code in [$\log_2$ N, N]. Furthermore, Miguer and Sergio

[11]proposed a new genetic operator to avoid invalid individuals and reduce the search space of the genetic algorithm.

Although there are already many papers discussing ECOC, the application of ECOC on microarray data is just at the beginning. Different from regular datasets, due to the small sample size of microarray data, a validation set is not affordable in the classification process, so it is much more complicated. In this paper, we propose a novel ensemble of ECOC(**E-ECOC**) system work by integrating different ECOC coding matrix with local difference maximum selection(**LDMS**) or global difference maximum selection(**GDMS**) strategies. And the experiments on some microarray datasets proves that our method is effective.

The rest of the paper is organized as follows: Section 2 overviews the background of ECOC framework. In Section 3, we present the E-ECOC framework. Section 4 is devoted to presenting the experimental results. Finally, Section 5 concludes the paper.

2 Error Correcting Output Codes

Let K denotes a set of unique labels, $K = \{k_1, k_2, \ldots, k_N\}$, where N means the number of classes (N > 2). Let S denotes a set of samples, $S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_L, y_L)\}$. X_i is the features vector represent the sample S_i, and y_i is the class label to which S_i belongs. Besides, $y_i \in K$. L means the number of samples. And Let D denotes a set of dichotomizers according to the ECOC coding matrix $D = \{d_1, d_2, \ldots, d_M\}$.

The basis of ECOC framework is building a unique "code word" for each class. The elements within the coding matrix of size M * N belong to the set {-1, +1} or {-1, 0, +1}. Each row represents a class, and there are M classes totally. Meanwhile, each column is interpreted as a binary classifier, and the original class label is re-calibrated into binary classes, which is named as meta-class. For instance, suppose a sample (X, y) belonging to class i. It will be re-labeled as positive in j-th dichotomizer when $ECOC(i, j)$ =1; otherwise (X, y) will be re-labeled as negative when $ECOC(i, j)$=-1. Moreover, (X, y) will be neglected when $ECOC(i, j)$=0.

(a) Encoding Algorithms

Encoding matrix plays an important role in ECOC framework, because it describes how to decompose a multi-class classification into a set of binary problem. In [12], the researchers summarized the methods to build coding matrix into two categories: static method and dynamic method. The static method commonly constructs coding matrix independent of base classifiers and datasets. There are four kinds of static coding design schemes, including One Vs. One, One Vs. All, dense random, and sparse random. Dynamic methods construct problem-dependent encoding matrixes, so they are more flexible comparing with static schemes.

The researchers take two factors into consideration: row separation/column separation and matrix validity. Row separation refers to the distance between any pair of code words, and column separation indicates the difference degree within each binary classifier pair. Both should be as large as possible, so as to reduce the correlation between base classifiers.

The coding matrix may not be correctly constructed, and there are some essential rules to check the legality of the matrix, as shown in equation 1-4[13]. Equation 1 indicates that each column of the encoding matrix comprises at least one +1 and -1. AHD represents attenuated Hamming distance. Equation 2 shows that the minimum Hamming distance between two rows should be at least one, which means all 0s, all +1s and all -1s are not correct. Equation 3 means if there is converse relationship between any two rows, the encoding matrix is invalid. Equation 4 indicates that the number of binary classifiers should be at least $\log_2 M$. Validity checking can provide pseudo integrity protection while constructing coding matrix.

$$\min\left(\delta AHD\left(r^i, r^l\right)\right) \geq 1, \forall i,k : i \neq k \neq, i,k \in [1,\ldots,N]. \tag{1}$$

$$\min\left(\delta HD\left(d^j, d^l\right)\right) \geq 1, \forall j,l : j \neq l,\ j,l \in [1,\ldots,M]. \tag{2}$$

$$\min\left(\delta HD\left(d^j, -d^l\right)\right) \geq 1, \forall j,l : j \neq l,\ j,l \in [1,\ldots,M]. \tag{3}$$

$$N \geq \log_2 M \tag{4}$$

(b) Decoding Strategies

When testing an unlabeled sample X*, each binary classifier gives an output, and the group of outputs makes up a vector V* with length L. Then, the distance between the output vector and code words within the coding matrix is calculated, and the code word with the minimum distance will be the class label to which X* belongs. The procedure is called decoding. There are different decoding strategies. Among them, hamming decoding is the most commonly used, as is shown in equation 5-6. It has obvious drawbacks, because it requires each binary classifier produces hard outputs, +1 or -1. With Euclidean decoding strategy, this problem can be solved, and the output of each classifier could be the confidence to positive class or negative class as shown in equation 7.

$$\mathrm{HD}\left(\mathrm{V}^*, \mathrm{y_i}\right) = \sum_{j=1}^{n} \frac{\left(1 - sign\left(\mathrm{V}^{*j} \times y_i^j\right)\right)}{2}. \tag{5}$$

$$\mathrm{y} = \min_{i=\{1,\ldots,n\}} \mathrm{HD}\left(\mathrm{V}^*, \mathrm{y_i}\right). \tag{6}$$

$$\mathrm{ED}\left(\mathrm{V}^{*}, y_i\right)=\sqrt{\sum_{j=1}^{n}\left(\mathrm{V}^{*j}-y_i^j\right)^2}\ . \tag{7}$$

Besides distance based decoding strategies, researchers also proposed some other schemes based on loss function[4]. The loss function is calculated firstly according to the output vector V* as shown in equation 8. Loss function $L\left(\theta\right)$ depends on the characteristics of the base classifier, and the most commonly used functions are $L\left(\theta\right)=-\theta$ (LLD) and $L\left(\theta\right)=e^{-\theta}$ (ELD). Then, the code word with the minimum loss function value is picked as the class label for a sample. . Moreover, decoding strategies based on probability have been proposed, which take probability estimation and confidence into consideration.

$$\mathrm{LB}\left(\mathrm{V}^{*}, y_i\right)=\sum_{j=1}^{n}L\left(\mathrm{V}^{*j}\times y_i^j\right). \tag{8}$$

3 Ensemble of ECOC

The most important purpose while designing ECOC coding matrix is to improve the error correction capability of the matrix. According to the theory of error-correcting, the matrix could fix d bits' error if the code matrix's minimum hamming distance equals 2d + 1. Therefore, many random-based algorithms and data-dependent algorithms try to maximize the minimum hamming distance. However, the ability to detect and correct errors depends on whether the errors occur independently. In the ECOC framework, the efforts to improve the binary classifiers' mutual independence is reasonable and essential. In [12], researchers uses different feature subsets for each dichotomizer, leading to more independent classifiers.

We design a new method to ensemble ECOC coding matrixes called ***E-ECOC***. The strategy consists of designing multiple ECOC coding matrixes, and then ensemble the matrices with high diversity. That is, a multi-class problem is solved by a set of different ECOC coding matrixes consequently, so as to increase the overall system accuracy. And the coding matrices are produced based on different problem dependent algorithms, and different feature subsets are used to construct the matrix within one same algorithm. The notation used to measure the difference is show in equation 9. We design two strategies to ensemble ECOC coding matrixes, the first one called as Local Difference Maximum Selection (L-DMS). Different feature subsets to construct the problem dependent coding matrix, and the top coding matrixes are chosen to solve the original multi-class problem. The second is called as Global Difference Maximum Selection (G-DMS). Different algorithms are applied to construct coding matrixes, and for each algorithm, different feature subsets are used. Then, we calculate the global difference degree and choose top coding matrixes with

maximum difference degree. The process of choosing coding matrices is shown in Figure 1.

$$Diff\left(E\left(M,N^{*}\right),E\left(M,N\right)\right)=\sum_{i=1}^{N^{*}}\min_{j=1}^{N}\left(L\left(d_{i},d_{j}\right)\right) \tag{9}$$

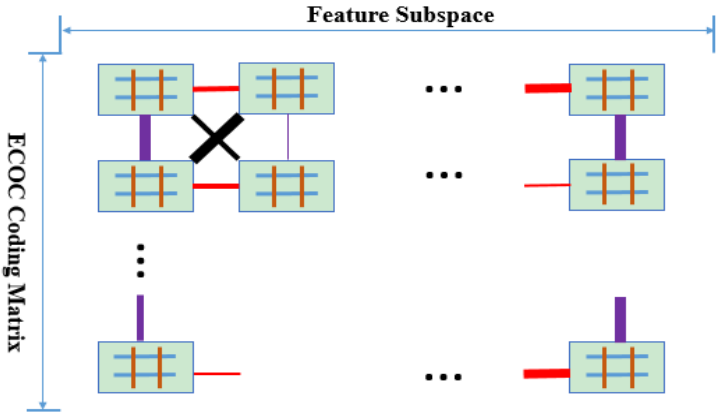

Fig. 1. ECOC coding matrixes' mutual difference degree, basis for ECOC ensemble. Red lines mean the local diversity among one same ECOC algorithm with different feature subsets (L-DMS). Purple lines and dark lines including the red lines indicate the global diversity among different ECOC coding matrixes with different feature subsets (G-DMS). The thickness of the line illustrates the difference degree.

4 Experiments and Analysis

ECOC library [14]is used to implement the ECOC algorithm framework, and three ECOC methods are used: DECOC[6], forest-ECOC[15], and ECOC-One[16]. The decoding method uses the default Hamming distance function. Two kinds of base classifiers are applied: KNN (k=3), and SVM (Lib-SVM library[17]). Other parameters use the default settings. The feature selection methods include Su[18], Laplacian Score[19] and t-test[20]. Moreover, the feature size within a same ECOC coding matrix increases from 20 to 200, and the step size is 20. We apply E-ECOC method to two well-known cancer datasets: Cancers [21], and Breast cancer dataset [22]. Table 1 shows the performance for each single ECOC coding matrix, and Table 1, 2 summarizes the ensemble results. Methods (a), (b), and (c) mean one single ECOC coding matrix with feature selection. Methods (d), (e), and (f) select from one same encoding algorithm with different feature subsets, which is called as L-DMS. Method (g) selects ECOC coding matrixes from different encoding algorithms and each constructed with different features, which is named as G-DMS.

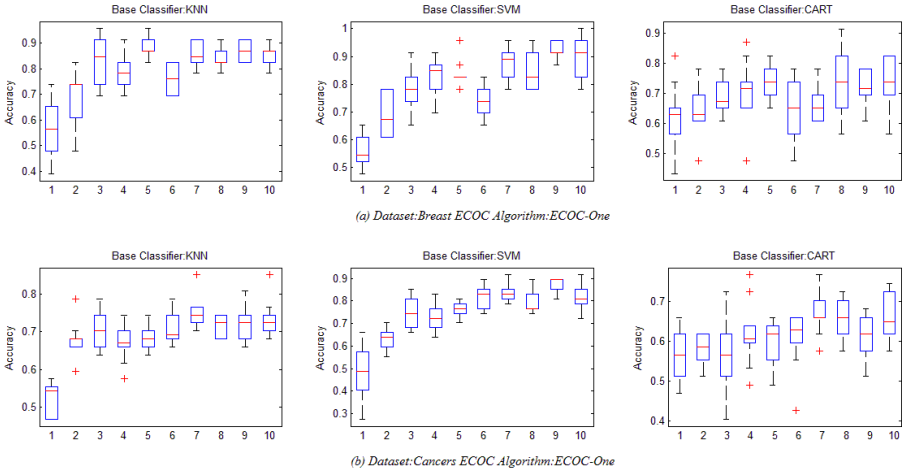

Fig. 2. Typical results of classification accuracy obtained by individual ECOC-One with different number of genes selected by Laplacian Score: (a) Dataset: Breast (b) Dataset: Cancers

Table 1. The comparison of average and best classification accuracies among individual ECOC coding matrixes and different ECOC ensemble method for dataset Breast

Breast		Base Classifier: KNN					
		Su		Laplacian Score		t-test	
		A_{avg}	A_{best}	A_{avg}	A_{best}	A_{avg}	A_{best}
(a)ECOC-One		56.52±1.09	73.91	69.13±1.07	82.61	69.13±0.35	78.26
(b)DECOC		65.65±0.61	78.26	72.61±0.17	78.26	62.17±0.51	73.91
(c)Forest-ECOC		63.04±0.47	73.91	76.52±0.39	82.61	64.35±1.42	82.61
L-DMS	(d)ECOC-One	73.04±0.79	**86.96**	82.61±0.67	**100.00**	**92.61±0.21**	**100.00**
	(e)DECOC	75.22±0.21	82.61	84.78±0.26	91.30	89.13±0.22	95.65
	(f)Forest-ECOC	73.91±0.42	82.61	**89.13±0.30**	95.65	89.57±0.30	95.65
(g)G-DMS		**78.26±0.42**	**86.96**	83.91±0.21	91.30	87.39±0.27	95.65
Breast		Base Classifier: SVM					
		Su		Laplacian Score		t-test	
		A_{avg}	A_{best}	A_{avg}	A_{best}	A_{avg}	A_{best}
(a)ECOC-One		55.65±0.29	65.22	68.26±0.59	78.26	63.04±1.10	82.61
(b)DECOC		63.91±0.25	69.57	66.96±0.39	73.91	59.13±0.93	73.91
(c)Forest-ECOC		68.26±0.30	73.91	78.26±0.34	86.96	64.35±1.08	82.61

Table 1. (*Continued.*)

L-DMS	(d)ECOC-One	86.96±0.08	91.30	**99.13±0.03**	**100.00**	**92.61±0.21**	**100.00**
	(e)DECOC	88.26±0.21	91.30	87.39±0.27	95.65	89.13±0.22	95.65
	(f)Forest-ECOC	**89.13±0.43**	**100.00**	86.52±0.27	91.30	89.57±0.30	95.65
(g)G-DMS		77.83±0.52	91.30	86.09±0.12	91.30	93.04±0.26	**100.00**

Table 2. The comparison of average and best classification accuracies among individual ECOC coding matrixes and different ECOC ensemble method for dataset Cancers

Cancers		Base Classifier: KNN					
		Su		Laplacian Score		t-test	
		A_{avg}	A_{best}	A_{avg}	A_{best}	A_{avg}	A_{best}
(a)ECOC-One		52.13±0.18	57.45	57.23±0.38	63.83	59.15±0.47	68.09
(b)DECOC		**67.87±0.23**	**78.72**	61.06±0.19	68.09	70.00±0.15	76.60
(c)Forest-ECOC		56.17±0.40	65.96	43.83±0.18	51.06	54.04±0.49	63.83
L-DMS	(d)ECOC-One	57.23±0.30	65.96	73.40±0.16	78.72	79.15±0.06	80.85
	(e)DECOC	63.62±0.31	72.34	74.47±0.10	78.72	**80.43±0.16**	**87.23**
	(f)Forest-ECOC	60.43±0.29	74.47	**76.17±0.43**	**89.36**	79.36±0.25	**87.23**
(g)G-DMS		65.96±0.35	74.47	74.04±0.15	80.85	78.09±0.18	82.98
Cancers		Base Classifier: SVM					
		Su		Laplacian Score		t-test	
		A_{avg}	A_{best}	A_{avg}	A_{best}	A_{avg}	A_{best}
(a)ECOC-One		48.30±1.49	65.96	46.81±0.33	55.32	50.21±0.63	65.96
(b)DECOC		63.19±0.22	70.21	52.34±0.11	59.57	61.70±0.46	76.60
(c)Forest-ECOC		57.87±0.14	61.70	59.15±0.24	48.94	58.51±0.44	65.96
L-DMS	(d)ECOC-One	81.49±0.11	87.23	86.17±0.25	93.62	87.23±0.27	**95.74**
	(e)DECOC	79.79±0.31	**89.36**	82.55±0.11	87.23	84.47±0.11	89.36
	(f)Forest-ECOC	78.51±0.26	85.11	**87.45±0.18**	**95.74**	**91.49±0.06**	**95.74**
(g)G-DMS		**83.40±0.11**	**89.36**	85.96±0.43	93.62	87.23±0.11	93.62

From Fig. 2, it can be found that the performance of the ECOC-One methods varies greatly with different feature subsets, which indicates the performance of data-dependent ECOC coding matrixes vary greatly. Comparing with individual ECOC coding matrixes, E-ECOC ensembles achieve better results. From Table. 1 and table. 2, E-ECOC with SVM generally has better average results. For dataset Breast, the best results reach 99.13±0.03. Its parameters include ECOC-One, SVM as base classifier

and t-test for feature selection. For dataset Cancers, the best results reach 91.49±0.06. Its parameters include forest-ECOC, SVM as base classifier and t-test for feature selection. Furthermore, L-DMS has similar performance comparing with G-DMS.

5 Conclusions

In this paper, we applied ECOC framework to tackle the microarray data classification problem. In this ensemble scheme, individual ECOC coding matrixes are selected according to the mutual diversity measures. Therefore, ECOC ensemble are used to solve the original multi-class classification problem. Two strategies including different feature subsets and different data-dependent ECOC coding matrixes are applied to promote diversity. The experimental results show that ECOC ensemble algorithm is an effective method for microarray classification, which usually leads to better accuracy. Furthermore, ECOC ensemble is more robust method comparing with individual ECOC.

References

1. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. arXiv preprint cs/9501101 (1995)
2. Escalera, S., Pujol, O., Radeva, P.: On the decoding process in ternary error-correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 120–134 (2010)
3. Kong, E.B., Dietterich, T.G.: Error-Correcting Output Coding Corrects Bias and Variance. In: ICML 1995 (1995)
4. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research 1, 113–141 (2001)
5. Masulli, F., Valentini, G.: Effectiveness of error correcting output codes in multiclass learning problems. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 107–116. Springer, Heidelberg (2000)
6. Pujol, O., Radeva, P., Vitria, J.: Discriminant ecoc: A heuristic method for application dependent design of error correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(6), 1007–1012 (2006)
7. Crammer, K., Singer, Y.: On the learnability and design of output codes for multiclass problems. Machine Learning 47(2-3), 201–233 (2002)
8. Bautista, M.Á., et al.: Minimal design of error-correcting output codes. Pattern Recognition Letters 33(6), 693–702 (2012)
9. García-Pedrajas, N., Fyfe, C.: Evolving output codes for multiclass problems. IEEE Transactions on Evolutionary Computation 12(1), 93–106 (2008)
10. Lorena, A.C., Carvalho, A.C.: Evolutionary design of multiclass support vector machines. Journal of Intelligent and Fuzzy Systems 18(5), 445–454 (2007)
11. Escalera, S., et al.: Subclass problem-dependent design for error-correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(6), 1041–1054 (2008)
12. Bagheri, M.A., Montazer, G.A., Kabir, E.: A subspace approach to error correcting output codes. Pattern Recognition Letters 34(2), 176–184 (2013)

13. Bautista, M.Á., et al.: On the design of an ECOC-Compliant Genetic Algorithm. Pattern Recognition 47(2), 865–884 (2014)
14. Escalera, S., Pujol, O., Radeva, P.: Error-Correcting Ouput Codes Library. J. Mach. Learn. Res. 11, 661–664 (2010)
15. Escalera, S., Pujol, O., Radeva, P.: Boosted Landmarks of Contextual Descriptors and Forest-ECOC: A novel framework to detect and classify objects in cluttered scenes. Pattern Recognition Letters 28(13), 1759–1768 (2007)
16. Escalera, S., Pujol, O., Radeva, P.: ECOC-ONE: A novel coding and decoding strategy. In: 18th International Conference on Pattern Recognition, ICPR 2006. IEEE (2006)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
18. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5, 1205–1224 (2004)
19. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS (2005)
20. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3(2), 185–205 (2005)
21. Su, A.I., et al.: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Research 61(20), 7388–7393 (2001)
22. Perou, C.M., et al.: Molecular portraits of human breast tumours. Nature 406(6797), 747–752 (2000)