Proceedings of the 2014 International Conference on Machine Learning and Cybernetics, Lanzhou, 13-16 July, 2014

AN INDICATOR-BASED SELECTION MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM WITH PREFERENCE FOR MULTI-CLASS ENSEMBLE

JING-JING CAO[1], SAM KWONG[2], RAN WANG[2,3], KE LI[2]

[1] School of Logistics Engineering, Wuhan University of Technology, Wuhan 430070, China
[2] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518067, China
E-MAIL: bettymore0501@gmail.com, cssamk@cityu.edu.hk

Abstract:

One of the most difficult components for multi-class classification system is to find an appropriate error-correcting output codes (ECOC) matrix, which is used to decompose the multi-class problem into several binary class problems. In this paper, an indicator based multi-objective evolutionary algorithm with preference involved is designed to search the high-quality ECOC matrix. Specifically, the Harrington's one-sided desirability function is integrated into an indicator-based evolutionary algorithm (IBEA), which aims to approximate the relevant regions of pareto front (PF) according to the preference of the decision maker. Simulation results show that the proposed approach has better classification performance than compared multi-class based algorithms

Keywords:

Error-correcting output coding; Indicator-based evolutionary algorithm; Harrington's one-sided desirability function; Pareto front; Multi-class problem

1. Introduction

Ensemble methods have been proven to be powerful to enhance performance of a single classifier by means of combing weak learners. Common techniques of base classifier for ensemble are originally designed for two-class problems since the decision boundary of binary classifier is simple and easy to be distinguished. Such kinds of algorithms include support vector machine (SVM) [1], Decision stump [2] and so on. However, many real-world applications have more than two classes and they are considered as multi-class problems. The multi-class problem is more complex than a binary class problem since the instance with multi-class has much higher probability to be misclassified than the instance with two-class. Thus, investigating multi-class problem becomes an active research area.

Most of the researchers decompose a multi-class problem into many two-class problems. There are three types of popular techniques: one-vs-one (OVO) [3], one-vs-all (OVA) [4] and error-correcting output codes (ECOC) [5]. One-vs-one (OVO) method selects a pair of classes for constructing each base classifier, by which it splits the c class problem into $c(c-1)/2$ binary problems. One-vs-all (OVA) method is considered as an efficient approach by learning a classifier between one class and the remaining classes. ECOC [5] is developed under the framework of code matrix $M \in \{-1, 1\}$, where 1 indicates the positive instance label and -1 corresponds to the negative instance label of a binary classifier. The matrix has a size of $k \times l$ with each column represents the output label of binary classifier and each row is designated as the codeword of each class. In fact, it can be induced that OVO and OVA as a special case of ECOC system. Further, searching for an optimal ECOC matrix is proven to be an NP hard problem [6, 7]. One effective way to find the best ECOC is to employ evolutionary algorithms (EAs), such as genetic algorithms (GAs). However, the existing problems for applying GAs into supervise learning problem rely on two folds: 1) Though the task for classification is to get the best classification accuracy, other qualities also should be considered for the goodness of ECOC matrix, such as row separation and column diversity of matrix. Applying single objective GAs (SOGAs) may not be good enough to balance the relationship among these criteria. 2) Multi-objective EAs (MOGAs) has obtained popularity since it aims at finding a trade-off among the objectives. In ECOC scenario, SPEA2 has been utilized to search the best matrix but the experimental results are not satisfactory[8]. Thus, how to devise a good multi-objective genetic algorithm becomes a challenge for ECOC problem.

978-1-4799-4215-2/14/$31.00 ©2014 IEEE

The main contribution of this paper is to include the Harrington's one-sided desirability function into the indicator-based selection evolutionary algorithm (IBEA), which aims to approximate the relevant regions of pareto front (PF) according to the preference of the decision maker. By doing so, all the objectives are mapped to the interval between 0 and 1. Three objectives are considered in this work: (1) training accuracy, (2) average training accuracy of the binary classifiers and (3) minimum relative hamming distances among codewords. Among these three objectives, the accuracy has been given a higher preference to guide the search along the region with higher accuracy rate.

2. Related Work

In recent years, many researchers have done several works to find the best ECOC, and GAs are regarded as efficient approaches among these methods. Alba et al. [7] design a local hybrid search approach named genetic algorithm with repulsion algorithm (GARA) for telecommunications systems, then this algorithm is adopted by Pimenta et al. [9] to generate the ECOC matrices. García et al. [10] design an evolutionary algorithm to search the "*best code matrix*" with each chromosome denotes a complete classifier. Lorena et al.[8] summarize several technologies on designing code-matrix for multi-class problems. They applied a multi-objective genetic algorithm: SPEA2 (Strength Pareto Evolutionary Algorithm 2) algorithm [11] to search the best code-matrix. However, the experimental results show that SPEA2 fails to find a better solution by compared with OVA with the same columns.

The concept of the hypervolume measure ($\mathcal{S}$-metric) is firstly proposed by Zitzler and Thiele [12]. Beume et al. [13] applies the $\mathcal{S}$-metric in an evolutionary multi-objective optimization algorithm (EMOA) as selection procedure. The $\mathcal{S}$-metric is considered as one of the most successful criteria for deriving a well-distributed set of approximation PF. Further, to evaluate the quality of the obtained approximation PF, the $\mathcal{S}$-metric is also regarded as an important quality indicator. As a matter of fact, the $\mathcal{S}$-metric is often designed for solving MOO problem and rarely employed in pattern recognition field since it focuses on finding the trade-off between convergence and diversity of Pareto-optimal set. However, for supervised learning problem, the desired classifier tends to predict the label of instances as accurate as possible. Thus, utilizing the preference based multi-objective GAs can give a compromise scheme to investigate the interest regions for supervised learning task. In addition, the IBEA used in this paper mainly employs a binary performance measure (indicator) in the environmental selection process and uses it as the optimization goal.

3. Preliminary Knowledge

3.1. Dense Decoding Methods

There are two steps can be implemented in ECOC system: the coding step and the decoding step. As mentioned in the introduction of this paper, the coding step applies the decomposition strategies to partitioning the multi-class labels into several bipartitions. Whereas the decoding phase consist of selecting the class (row of ECOC matrix) with the least distance between testing codeword and base codeword of each class. The codeword for a testing instance is generated by the output of each binary classifier.

Suppose we have a ECOC matrix $M_{k\times T}$ with values of $\{-1,+1\}$, the decision vector $f(x)$ of instance x is given as follows:

$$f(x) = \arg\min_{c=1,\ldots,k} DM(\boldsymbol{x}, M_{(c,\cdot)}) \tag{1}$$

Typically, the decoding methods for dense code is mainly given as follows [14]:

• Hamming decoding This technique is the most frequently used method for ECOC decoding, which is under the assumption that the learning task is regarded as an error-correcting communication. The hamming distance between the output of the t_th binary classifier and each row of M is given as follows:

$$DM_{HD}(\boldsymbol{x}, M_{(c,\cdot)}) = \sum_{t=1}^{T} \frac{1 - sign(f_t(x)\cdot M_{(c,t)})}{2} \tag{2}$$

• Euclidean decoding[15] This technique is also a common method by directly employ the Euclidean distance.

$$DM_{ED}(x, M_{(c,\cdot)}) = \sqrt{\sum_{t=1}^{T}(f_t(x) - M_{(c,t)})^2} \tag{3}$$

In the framework of decoding strategies, it has been shown that the Euclidean decoding is in proportion to the hamming decoding in dense coding case [14]. This case is only related to the number of failures between two codewords.

3.2. Design Principle of Multi-Class ECOC Method

In general, evaluation of the goodness of ECOC matrix depends on many folds. In GA learning, these criteria can be used as objectives to guide selection procedure.

• Lacc: Training accuracy of the individual k in the population:

$$Lacc_k^t = \frac{1}{N}\sum_{n=1}^{N} I(f(x_n) = y_n), \quad (4)$$

where t denotes the current iteration, $f(x_n)$ is defined in Eq. (1) and $I(\cdot)$ equals to 1 if $f(\boldsymbol{x}_n) = y_n$, otherwise, 0.

• BCacc: Average training accuracy of the binary classifiers for individual k:

$$BCacc_k^t = \frac{1}{Nl}\sum_{n=1}^{N}\sum_{p=1}^{l} I(f(\boldsymbol{x}_n, M_p) = y_n), \quad (5)$$

where $f(x_n, M_p)$ represents the output class label of pattern x_n, which is obtained by the p_th binary classifier and l is the length of the code.

• MHD: Minimum relative hamming distances among codewords. It is related to the "row separation" of the ECOC matrix; that is, codewords should be well-separated in hamming distance.

$$MHD_k^t = \min_{\substack{i,j=1,\ldots,k,\\ i\neq j}} \frac{DM(M_{(c_i,\cdot)}, M_{(c_j,\cdot)})}{l} \quad (6)$$

• CL: The length of codewords. This value should be minimized in order to decrease the redundancy of the codewords. Since each column of the matrix represents a binary classifier, the minimization of CL is equivalence of pruning the number of base binary learners.

From the perspective of communication techniques, MHD and CL are two criteria that conflicted with each other. Concretely, the larger value of MHD implies the higher error-correcting capacity, while the smaller value of CL indicates the faster transmission of error-correcting communication.

In general, the margin of the classification and the binary classifier's diversity measure can be considered as objectives as well. However, the practical behaviors of these two criteria are shown less attractive than $Lacc$, $BCacc$ and MHD criteria [10]. Thus in this work, we only apply $Lacc$, $BCacc$ and MHD as the objectives. Among these three objectives, the values of $Lacc$ and $BCacc$ are distributed in similar range, typically the $BCacc$ is a bit higher than $Lacc$, while for MHD, the value of individual is usually very small. Take the UCI data set "mfeat_mor" as an example, we randomly generate 100 ECOC matrices and evaluate their corresponding three objective values and ranges. This procedure is repeated ten times and the average range of each objective is obtained. For this dataset, the ranges of $Lacc$, $BCacc$ and MHD are $[0.7056, 0.7469]$, $[0.8497, 0.8874]$ and $[0.0244, 0.0412]$. From these ranges, it can be seen that MHD has a different range of $Lacc$ and $BCacc$. Though the range of each objective value does not affect the $I(f_i(x_1), f_i(x_2)), i = 1, 2, 3$, we want to limit all the value in a similar range for illustrating the experimental analysis more clearly. Hence, we add a term C^t in the form of evaluating the value of MHD in each iteration. Further, it should be noted that the criteria $Lacc$, $BCacc$ and MHD are the objectives that needed to be maximized, then the following objectives are employed to be minimized in this work:

$$\begin{aligned} f(x_k) &= [f_1(x_k), f_2(x_k), f_3(x_k)]^{\mathcal{T}} \\ &= [1 - Lacc_k^t, 1 - BCacc_k^t, 1 - MHD_k^t - C^t]^{\mathcal{T}}, \end{aligned} \quad (7)$$

where $C^t = \min(\sqrt{\frac{\overline{Lacc^t}^2 + \overline{BCacc^t}^2}{2}}, 1 - \max_k MHD_k^t)$, and $\overline{Lacc^t}$ and $\overline{BCacc^t}$ denote the average $Lacc$ and $BCacc$ values among all the individuals in the current population, respectively. $(1 - \max_k MHD_k^t)$ is used to make sure that $1 - MHD_k^t - C^t \geq 0$.

3.3. Desirability Function

The hypervolume based MOEA mainly focus on generating the solution with regards to two factors: one is the convergence of the non-dominated solutions, while the other is a well-distribution of those solutions. However, in some applications, the decision maker only pursues to the relevant regions of the pareto front (PF). The desirability functions (DFs) are proposed to map each objective to the interval [0, 1] and different parameter settings of DFs show the preference of different objectives. Usually, the DFs can transform each objective into a nonlinear shaped curve. In the framework of multiobjective industrial quality control, Harrington [16] introduces the definition of desirability. In fact, any function that can map the objective space to domain [0, 1] can be considered as a desirability. The following equation is a well-known one-sided Harrionton DF with the form of the Gompertz-curve [17]:

$$d(Y) = exp(-exp(-b_0 + b_1 Y)), \quad (8)$$

where b_0 and b_1 are the parameters that determined by the following formula:

$$\begin{aligned} b_0 &= -log(-log(d^{(1)}) - b_1 Y^{(1)}), \\ b_1 &= (-log(-log(d^{(2)})) + log(-log(d^{(1)})))/(Y^{(2)} - Y^{(1)}) \end{aligned} \quad (9)$$

In the above equations, $(Y^{(1)}, d^{(1)})$ and $(Y^{(2)}, d^{(2)})$ represent two objective values and the corresponding desirability. In Tobias and Heike's work [18], they point out that the setting values of $d^{(1)}$ and $d^{(2)}$ in industrial application is approximately 0.99 and 0.01 respectively. One of the advantages of this DF is that it maintains the dominance relations of individuals in each objective dimension due to its monotonicity property. Another merit is that the generated desirability is on the same scale after DF mapping, which avoids the bias range of the solution in approximated PF.

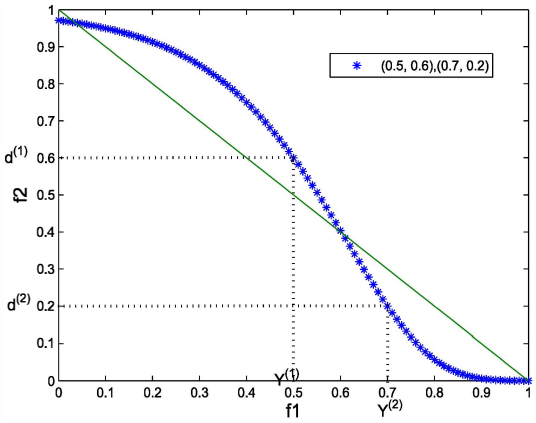

Figure 1. Shape of the Gompertz-curve with objective values of $(Y^{(1)}, d^{(1)}) = (0.5, 0.6)$ and $(Y^{(2)}, d^{(2)}) = (0.7, 0.2)$.

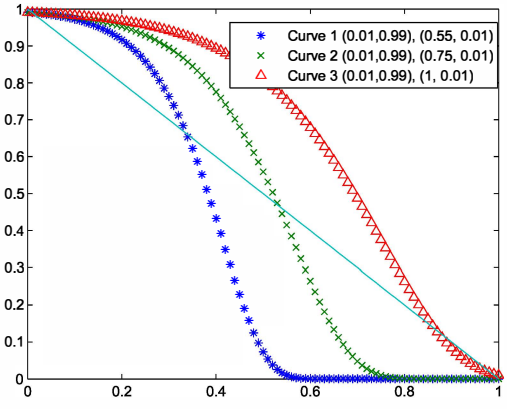

Figure 2. Shape of the Gompertz-curve with objective values of $(Y^{(1)}, d^{(1)})$ and $(Y^{(2)}, d^{(2)})$.

An example of the Gompertz-curve shape is shown in Fig.8(a). It transforms the shape of PF ($f_2 = 1 - f_1$) by selecting parameter values as $(Y^{(1)}, d^{(1)}) = (0.5, 0.6)$ and $(Y^{(2)}, d^{(2)}) = (0.7, 0.2)$. This type of DF is a one-sided specification curve, which is designed for maximizing or minimizing the objectives. Fig. 1 shows three different transformation shapes of the PF in desirability space. The main goal for classification is to improve the learning accuracy in prediction phase. Regardless of the overfitting factor, we assume that the higher training accuracy probably implies the higher testing accuracy. Further, it is obvious that in ensemble theory, the classifier with high or average performance is much preferred. Based on the above discussions, the objective *Lacc* is selected to be transformed by one-sided Harrionton DF with parameter setting as $(Y^{(1)}, d^{(1)}) = (0.01, 0.99)$ and $(Y^{(2)}, d^{(2)}) = (1, 0.01)$, i.e., curve 3. Through this transformation, it can be seen from Fig. 8.(b) that the individual with small value in term of training accuracy would probably be assigned less value. For the individuals with high or medium value of training accuracy, they would be given larger values. It implies that the classifier with a higher capability to classify the training data is probably preferred during iterations. Compared with curve 3, curve 1 focuses on transforming medium values to small values, and curve 2 controls the medium values in a large range of change.

3.4. IBEA-DF Algorithm

According to the above analysis, we propose an indicator based multi-objective evolutionary algorithm with preference information included (IBEA-DF). The algorithm is designed base on the framework of IBEA [19] and step 2 and step 3 are added in IBEA-DF algorithm. The detailed algorithm is described in algorithm 1.

4. Experiments

In this section, we conduct the experiment on 10 multi-class data sets from the UCI Machine Learning Repository [20]. Table 1 represents a summary of these datasets, which are the benchmark problems consist at least six classes. LibSVM [21] is also used in both experiments, and SVM with gaussian RBF kernel ($e^{\gamma\|u-v\|^2}$) is adopted for all the methods. Specifically, the parameter γ is set as: $\gamma = 2^{-13}$. Regarding the evolutionary algorithm, the population size β and the maximum number of generations T are both set as 100. The crossover and mutation rate equal to 0.9 and 0.1. $\kappa = 0.5$ is used in fitness evaluation and the code length of each ECOC matrix is fixed as 50.

The experiment is designed by comparing the conventional one-vs-all, CHC [10], IBEA [19] and IBEA-DF algorithms.

Algorithm 1: IBEA_DF

Input: *Population size* β, *maximum number of generations* T, *fitness scaling factor* κ

Output: *Pareto set approximation* A

Step 1: Initialization[19]: Generate an initial population P of size β; set $t = 1$.

Step 2: Multi-objective values evaluation Calculate multi-objective values of individuals in P, i.e., $Lacc_k^t$, $BCacc_k^t$ and MHD_k^t. Then obtain $f(x_k) = [1 - Lacc_k^t, 1 - BCacc_k^t, 1 - MHD_k^t - C^t]^T$

Step 3: DF based preference transformation Define $Y = 1 - Lacc_k^t$, and transform Y by one-sided Harrionton DF as follows: $d(Y) = exp(-exp(-b_0 + b_1 \times Y))$,

Step 4: Fitness assignment[19]: Calculate fitness values of individuals in P, i.e., for all $x^k \in P$ set $F(f(x_k)) = \sum_{f(x_l) \in P \setminus \{f(x_k)\}} -e^{-I(\{f(x_l)\},\{f(x_k)\})/\kappa}$,

Step 5: Environmental selection: Repeat the steps 6-8 until the size of population P does not exceed β as original IBEA Algorithm [19] describes,

Step 6: Termination: If $t \geq T$ is satisfied then set A to the set of decision vectors represented by the non-dominated individuals in P. Stop.

Step 7: Mating selection: To fill the temporary mating pool P', binary tournament selection with replacement is applied on population P .

Step 8: Variation: Utilize recombination and mutation operators of CHC algorithm [10] to the mating pool P' and put the offspring into P. Let $t = t + 1$ and go to Step 2.

Table 1. DATASET INFORMATION

Data	Pattern #	Feature #	Class #
Chart	600	60	6
Dermatology	366	34	6
Ecoli	336	7	8
Libras	360	90	15
Mfeat_fac	2000	216	10
Mfeat_fou	2000	76	10
Mfeat_kar	2000	64	10
Mfeat_mor	2000	6	10
Mfeat_zer	2000	47	10
Yeast	1484	8	10

Table 2. AVERAGE OF TESTING ERROR BY COMPARING OVA, CHC, IBEA AND IBEA_DF ALGORITHMS WITH GAUSSIAN RBF KERNEL

Data	OVA	CHC	IBEA	IBEA-DF
Chart	.1037 ± .0093	.0233 ± .0059	.0200 ± .0058	.0190 ± .0037
Dermatology	.2967 ± .0133	.1399 ± .0089	.1388 ± .0108	.1235 ± .0105
Ecoli	.1786 ± .0111	.1482 ± .0110	.1482 ± .0152	.1429 ± .0060
Libras	.5406 ± .0325	.3056 ± .0276	.2922 ± .0242	.3017 ± .0275
Mfeat_fac	.8416 ± .0019	.3631 ± .0461	.3533 ± .0576	.3730 ± .0355
Mfeat_fou	.6472 ± .0034	.2737 ± .0198	.2530 ± .0194	.2371 ± .0201
Mfeat_kar	.2698 ± .0040	.0415 ± .0039	.0371 ± .0032	.0401 ± .0035
Mfeat_mor	.4084 ± .0398	.2877 ± .0043	.2827 ± .0082	.2873 ± .0032
Mfeat_zer	.2610 ± .0054	.1800 ± .0080	.1792 ± .0082	.1807 ± .0068
Yeast	.6340 ± .0116	.4236 ± .0074	.4340 ± .013	.4278 ± .0077
Average	.5151 ± .0110	.3489 ± .0119	.3449 ± .0138	.3444 ± .0104

IBEA algorithm remove the step 3 from IBEA-DF to show the performance of indicator based evolutionary algorithm. From Table 2, IBEA and IBEA-DF represent similar behavior, while on the average term, the performance of IBEA-DF is slightly better than IBEA algorithm. As for OVA and CHC methods, they only achieve the best performance on Mfeat_kar and Yeast data set, respectively. The results for CHC algorithm are worst but it cost the least computational time. Overall, the proposed indicator based algorithms IBEA and IBEA-DF perform better then traditional CHC and OVA algorithms.

5. Conclusions

In this paper, we have studied different ways to integrate multi-objective evolutionary algorithm into searching best E-COC matrix for multi-class recognition problem. We firstly apply indicator-based selection multi-objective evolutionary algorithm (IBEA) to replace single-objective based genetic algorithm since ECOC based multi-class problem need to be solved by considering more than one criterion. Further, a one-sided desirability function is utilized to assign preference to the training accuracy, which is considered as an important objective in genetic learning process. Accordingly, we have conduct experiment to analyze and show the performance of the proposed algorithms. There are still some more work could be done in the future. The first is to compare the differences among other multi-objective based genetic algorithms, such as NSGA-II. Another is to utilize different desirability functions or parameters to find the most appropriate algorithm for solving multi-class problem.

6. Acknowledgements

This work was partly supported by the Fundamental Research Funds for the Central Universities (WUT:2014-IV-054), National Natural Science Foundation of China under the Grant No. 61175123, and Shenzhen New Industry Development Fund under grant No. JCYJ20120617120716224.

References

[1] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000.

[2] W. Iba and P. Langley, "Induction of one-level decision trees," in *Proceedings of the Ninth International Conference on Machine Learning*, pp. 233–240, 1992.

[3] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.

[4] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[5] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Arxiv preprint cs/9501101*, 1995.

[6] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Machine Learning*, vol. 47, no. 2, pp. 201–233, 2002.

[7] E. Alba and J. Chicano, "Solving the error correcting code problem with parallel hybrid heuristics," in *Proceedings of the 2004 ACM symposium on Applied computing*, pp. 985–989, ACM, 2004.

[8] A. Lorena and A. Carvalho, "Evolutionary design of code-matrices for multiclass problems," *Soft Computing for Knowledge Discovery and Data Mining*, pp. 153–184, 2008.

[9] E. Pimenta and J. Gama, "A study on error correcting output codes," in *Artificial intelligence, 2005. epia 2005. portuguese conference on*, pp. 218–223, IEEE, 2005.

[10] N. García-Pedrajas and C. Fyfe, "Evolving output codes for multiclass problems," *Evolutionary Computation, IEEE Transactions on*, vol. 12, no. 1, pp. 93–106, 2008.

[11] E. Zitzler, M. Laumanns, L. Thiele, E. Zitzler, E. Zitzler, L. Thiele, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm," 2001.

[12] E. Zitzler and L. Thiele, "Multiobjective optimization using evolutionary algorithms-a comparative case study," in *Parallel Problem Solving from Nature-PPSN V*, pp. 292–301, Springer, 1998.

[13] N. Beume, B. Naujoks, and M. Emmerich, "Sms-emoa: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.

[14] S. Escalera, O. Pujol, and P. Radeva, "On the decoding process in ternary error-correcting output codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 120–134, 2010.

[15] T. Windeatt and R. Ghaderi, "Coding and decoding strategies for multi-class learning problems," *Information Fusion*, vol. 4, no. 1, pp. 11–21, 2003.

[16] E. Harrington, "The desirability function," *Industrial quality control*, vol. 21, no. 10, pp. 494–498, 1965.

[17] T. Wheldon, *Mathematical models in cancer research*. Hilger, 1988.

[18] T. Wagner and H. Trautmann, "Integration of preferences in hypervolume-based multiobjective evolutionary algorithms by means of desirability functions," *Evolutionary Computation, IEEE Transactions on*, vol. 14, no. 5, pp. 688–701, 2010.

[19] E. Zitzler and S. Künzli, "Indicator-based selection in multiobjective search," in *PPSN*, pp. 832–842, 2004.

[20] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[21] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.