Supplementary Material for "Evolutionary Classifier Chain for Multi-Dimensional Classification"

1 Example Diagram of the Encoding

To facilitate understanding, Figure S-I illustrates how a solution and a population are encoded.

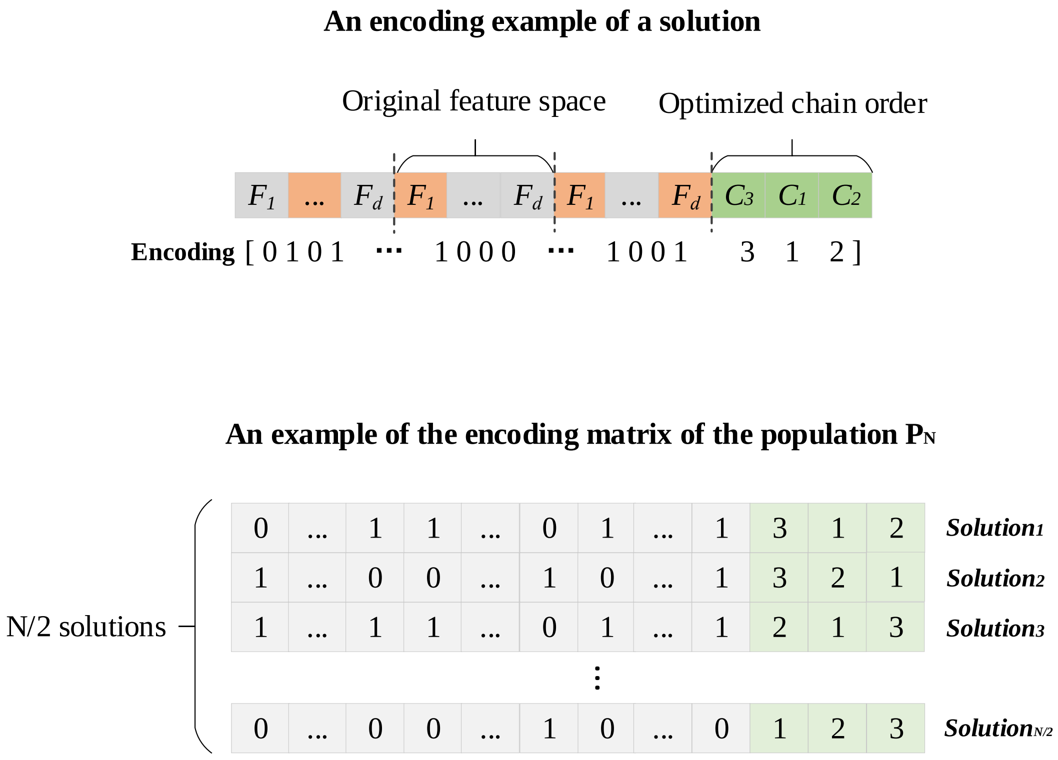

Figure S-I: Example diagram of the encoding of a solution and a population

2 Summary of the Notations

Table S-I summarizes the notations used to describe the ECCO approach, as well as other notations used in our paper.

Table S-I: Summary of the notations used in our paper.

Notation	Descriptions
d	number of features in original input space
q	number of class spaces (dimensions) in output space
K_j	number of class labels in the j-th class space $(1 \leq j \leq q)$
C_j	the j-th class space where $C_j = \{c_1^j, c_2^j, \ldots, c_{K_j}^j\}$ $(1 \leq j \leq q)$
c_a^j	the a-th class label in C_j $(1 \leq a \leq K_j)$
$\mathcal{X}$	the d-dimensional input (feature) space
$\mathcal{Y}$	the output space where $\mathcal{Y} = C_1 \times C_2 \times \ldots \times C_q$
N	the number of solutions contained in the overall population (the number of rows of the overall population matrix)
$proC$	the probability of crossover
$proM$	the probability of mutation
$\mathbf{P}_N$	the encoding matrix ($N/2 \times q(d+1)$ dimensional) of the parent of the dominance-based subpopulation
$\mathbf{P}_N(:, 1 : q * d)$	the encoding matrix ($N/2 \times q * d$ dimensional) of the part of input features in $\mathbf{P}_N$
$\mathbf{P}_M$	the encoding matrix ($N/2 \times q(d+1)$ dimensional) of the parent of the decomposition-based subpopulation
$\mathbf{P}_M(:, 1 : q * d)$	the encoding matrix ($N/2 \times q * d$ dimensional) of the part of input features in $\mathbf{P}_M$
$\leftarrow$ false	Assign 0 to the element in the corresponding position in the matrix.
$\mathbf{K}$	the logical matrix ($N/2 \times q * d$ dimensional) where each position of 1 represents the position of the element to be crossed
$\mathbf{S}$	the logical matrix ($N/2 \times q * d$ dimensional) where each position of 1 represents the position of the element to be mutated
$\mathbf{P}_1(\mathbf{K})$	Take the element of $\mathbf{P}_1$ in the corresponding position of 1 in the logical matrix $\mathbf{K}$.
$\mathbf{O}_1(\mathbf{K})$	Take the element of $\mathbf{O}_1$ in the corresponding position of 1 in the logical matrix $\mathbf{K}$.
$\mathbf{O}_N$	the encoding matrix ($N/2 \times q(d+1)$ dimensional) of the offspring of the dominance-based subpopulation
$\mathbf{O}_M$	the encoding matrix ($N/2 \times q(d+1)$ dimensional) of the offspring of the decomposition-based subpopulation
$\mathbf{O}_N = [\mathbf{O}_1; \mathbf{O}_2]$	Stack two matrices $\mathbf{O}_1$ and $\mathbf{O}_2$ into one whole matrix $\mathbf{O}_N$ by rows.
$\mathbf{O}_N(\mathbf{S}) = \sim \mathbf{O}_N(\mathbf{S})$	Inverse the elements of the matrix $\mathbf{O}_N$ that correspond to the positions in the logical matrix $\mathbf{S}$ where the element is 1.
$\mathbf{O}$	the overall offspring population matrix ($N \times q(d+1)$ dimensional) where each row represents the encoding of a solution
$\boldsymbol{x}_{Best}$	the encoding of the best solution on the Pareto front with the smallest f_2
$Pareto\ front$	the set of solutions where no objective can be improved without compromising at least one other objective (the front with the smallest number among the fronts assigned to all solutions)
f_1	One of the objectives to be optimized, i.e., $the\ ratio\ of\ selected\ features$ (calculated by the number of input features selected dividing $q * d$ for each solution)
f_2	One of the objectives to be optimized, i.e., $1 - Hamming\ Score$ (calculated by Eq. (4))
$\mathbf{P}_N \cup \mathbf{O}_N$	Combine all solutions in populations $\mathbf{P}_N$ and $\mathbf{O}_N$.
$\mathbf{P}_A$	the encoding matrix ($N \times q(d+1)$ dimensional) with all solutions for two subpopulations $\mathbf{P}_N$ and $\mathbf{P}_M$ together
$\boldsymbol{v}_{Index}$	the vector ($1 \times N$ dimensional) containing the numbers of the fronts where all the solutions in the population are located, where all the solutions are sorted by their numbers from smallest to largest
$\boldsymbol{a}$	the encoding solution in the population
$\boldsymbol{b}$	the encoding solution in the population
g	the aggregation function in the Tchebycheff method
$\boldsymbol{u}$	the encoding solution in the population
$\boldsymbol{\lambda}$	the two-dimensional weight vector with preferences for f_1 and f_2
Z^*	the ideal point
Ω	the set of all solutions in the population
$\mathcal{S}$	the test set
l	the ECCO predictive model: $\mathcal{X} \mapsto \mathcal{Y}$
$\boldsymbol{x_i}$	the feature vector of the i-th sample
$\boldsymbol{y_i}$	the class vector of the i-th sample
y_{ij}	the ground-truth labels for the j-th dimension of the i-th test sample
$\hat{y}_{ij}$	the predicted labels for the j-th dimension of the i-th test sample

3 Further Analysis for Performance Differences Across Datasets

In order to understand the specific reasons for the performance differences in the compared algorithms, we performed an analysis of the performance variations across datasets.

First, at the dataset level, the differences in #Dim and #Labels/Dim across different datasets directly affects the complexity of MDC problems. For example, the Oes97 dataset has 5 times more class dimensions than the Flare1 dataset. Therefore, all algorithms have significantly lower experimental metrics on the Oes97 dataset than on the Flare1 dataset. Second, the effect of optimization can vary across datasets with different strengths of class dependencies. For example, in datasets with strong class dependencies (e.g., Enb, WaterQuality, BeLaE), chain-order optimization may improve performance metrics more significantly. Third, the noise level and feature relevance of the dataset also affect the adaptation of the proposed algorithm ECCO. The performance of the metrics is improved after feature selection on some datasets for all class dimensions with significant redundant features. However, for datasets with strong interaction between features, the effect after feature selection may not be significant, especially if the number of features is small.

Overall, there are three main reasons for the performance differences of ECCO on different MDC datasets: the class and label dimensions of the dataset, the strength of class dependencies, and feature relevance.

4 Further Analysis of Ablation Experiments

In this section, we conduct experiments to verify the effectiveness of the two parts of the proposed chain order optimization and dimension-feature selection. For this purpose, two variants of the algorithms ECCO-RC and ECCO-AF are designed with random chain order and with all features, respectively. The results of the recorded experiments are shown in Tables S-II, S-III, S-IV and S-V. According to the results of Wilxcon rank-sum test and Friedman test, the performance of the algorithm decreases significantly when these two strategies are eliminated respectively. Therefore, these two strategies can improve the performance of the algorithm in solving the MDC problem. Better classification metrics are obtained when combining the two together. This is due to the fact that dimension-features are more discriminative than all features. Moreover, the optimized chain order is more consistent with the true class dependency.

Table S-II: Mean results of ECCO and two variant algorithms in terms of HS on seven representative datasets.

DataSets	HS		
	ECCO	ECCO-RC	ECCO-AF
Edm	**0.738**	0.734	0.688
Jura	**0.641**	0.563	0.622
Enb	**0.790**	0.783	0.782
WQanimals	**0.663**	0.662	0.662
BeLaE	**0.481**	0.468	0.473
Voice	**0.940**	0.935	0.906
Pain	**0.960**	0.959	**0.960**
Friedman's rank	**1.071**	2.500	2.429

Table S-III: Mean results of ECCO and two variant algorithms in terms of EM on seven representative datasets.

DataSets	EM		
	ECCO	ECCO-RC	ECCO-AF
Edm	**0.481**	0.469	0.438
Jura	**0.411**	0.306	0.405
Enb	**0.581**	0.565	0.565
WQanimals	**0.096**	0.093	0.093
BeLaE	**0.055**	0.054	0.044
Voice	**0.883**	0.871	0.816
Pain	**0.798**	0.795	0.796
Friedman's rank	**1.000**	2.429	2.571

Table S-IV: Mean results of ECCO and two variant algorithms in terms of SEM on seven representative datasets.

DataSets	SEM		
	ECCO	ECCO-RC	ECCO-AF
Edm	0.994	**1.000**	0.938
Jura	**0.872**	0.819	0.838
Enb	**1.000**	**1.000**	**1.000**
WQanimals	**0.281**	0.276	0.280
BeLaE	**0.196**	0.178	**0.196**
Voice	**0.997**	**0.997**	**0.997**
Pain	**0.896**	0.894	**0.896**
Friedman's rank	**1.571**	2.429	2.000

Table S-V: Summary of the Wilcoxon signed-ranks test for ECCO against its variants in terms of each evaluation metric at 0.05 significance level. The p-values are shown in the brackets.

ECCO against	HS	EM	SEM
ECCO-RC	**win**[1.22E-02]	**win**[1.04E-02]	**tie**[1.76E-01]
ECCO-AF	**win**[1.78E-02]	**win**[1.42E-02]	**tie**[1.08E-01]

5 Further Analysis of Parameter Sensitivity

To explore the sensitivity of the parameters used in this paper, we conduct experiments to verify the appropriateness of the population size settings. The performance metrics of ECCO and ECCO-P100 with population sizes of 200 and 100 on six representative datasets are recorded in Table S-VI. From the experimental results and the Wilcoxon signed-ranks test with significance of 0.05(p-value), it can be found that a population size of 200 has better convergence results. This setting provides a better trade-off between convergence and cost.

Table S-VI: Mean results of ECCO with different population size in terms of HS/EM/SEM on 6 datasets.

DataSets	HS		EM		SEM	
	ECCO	ECCO-P100	ECCO	ECCO-P100	ECCO	ECCO-P100
Jura	**0.641**	0.635	0.411	**0.419**	**0.872**	0.851
Enb	**0.790**	0.782	**0.581**	0.565	**1.000**	**1.000**
WQanimals	**0.663**	0.660	**0.096**	0.089	**0.281**	0.276
WQplants	**0.706**	**0.706**	0.172	**0.174**	**0.400**	0.399
WaterQuality	**0.692**	0.691	0.031	**0.033**	**0.110**	0.108
Voice	**0.940**	0.906	**0.883**	0.817	**0.997**	0.995
[p-value]	-	**win**[3.10E-02]	-	**tie**[3.89E-01]	-	**win**[2.56E-02]

6 Computational complexity analysis

The computational cost of the proposed method ECCO mainly comes from the nondominated sorting operation, environmental selection, and fitness evaluation. First, the computational complexity of the parent selection and reproduction operators is $O(N/2)$ in each subpopulation, respectively. N is the population size. Second, environmental selection is performed by a neighborhood-based Tchebycheff method in PM. Its computational complexity is $O(N/2 \cdot T \cdot M)$, where T and M are the neighborhood size and the number of objectives, respectively. Third, the computational complexity of the nondominated sorting operation in P_N and the merged population P_A are $O(M \cdot (N/2)^2)$ and $O(M \cdot N^2)$, respectively. The computational complexity of the crowding distance metric in P_N is $O(M \cdot N/2 \cdot \log(N/2))$. Finally, the maximum computational complexity in evaluating the fitness of each solution is $2 \cdot O(N/2 \cdot (D + S \cdot \mathrm{dim}))$. Among them, D is the total number of features, S is the number of samples in the training set, and dim is the number of class dimensions. Thus, the overall complexity of the proposed ECCO in each generation is: $2 \cdot 2 \cdot O(N/2) + O(N/2 \cdot T \cdot M) + O(M \cdot (N/2)^2) + O(M \cdot N/2 \cdot \log(N/2)) + O(M \cdot N^2) + 2 \cdot O(N/2 \cdot (D + S \cdot \mathrm{dim})) = \max\{O(M \cdot N^2), O(N \cdot (D + S \cdot \mathrm{dim}))\}$.