On the Optimization of Embedding Spaces via Information Granulation for Pattern Recognition

Alessio Martino, Fabio Massimo Frattale Mascioli, Antonello Rizzi
Department of Information Engineering, Electronics and Telecommunications
University of Rome "La Sapienza"
Via Eudossiana 18, 00184 Rome, Italy
{alessio.martino, fabiomassimo.frattalemascioli, antonello.rizzi}@uniroma1.it

***Abstract*—Embedding spaces are one of the mainstream approaches when dealing with structured data. Granular Computing, in the last decade, emerged as a powerful paradigm for the automatic synthesis of embedding spaces that, at the same time, yield an interpretable model on the top of meaningful entities known as "information granules". Usually, in these contexts, one aims at finding the smallest set of information granules in order to boost the model interpretability while keeping satisfactory performances. In this paper, we add a third objective, namely the structural complexity of the resulting model and we exploit three biology-related case studies related to metabolic networks and protein networks in order to investigate the link between classification performances, embedding space dimensionality and structural complexity of the resulting model.**

***Index Terms*—Granular Computing, Embedding Spaces, Support Vector Machine, Systems Biology, Topological Data Analysis, Computational Biology.**

I. Introduction

Many interesting, yet challenging, problems in pattern recognition deal with structured data such as images, videos, graphs and sequences. These examples of structured data are commonly used for modelling several real-world systems: think about sequence matching in computational biology, where RNA, DNA and proteins are commonly described by sequences of nucleotides or amino-acids; (cyber)security tasks likely involve images and/or videos captured by CCTVs; in social networks and systems biology, graphs are able to model relationships between users (the former) or interacting atomic elements (the latter).

Pattern recognition problems defined in structured domains are usually featured by complex decision functions and traditional techniques are likely to fail. A common approach consists in mapping the structured domain $\mathcal{X}$ towards $\mathbb{R}^n$: that is because the umbrella of soft computing techniques are well-established when it comes to process input patterns lying in a geometric input space such as the Euclidean one. However, designing the mapping function is not trivial. A naïve approach consists in extracting numerical features to be concatenated in a vector form [1]–[3]. Despite its straightforwardness, this approach can easily destroy valuable information useful for the pattern recognition problem at hand due to the intrinsic high compression in feature generation. For this approach to be successful, a deep knowledge on both the data and the problem at hand (i.e., on the underlying process to be modelled) is required for selecting a suitable subset of numerical features. Otherwise, trial-and-error steps are mandatory in order to find a suitable pattern representation which can, however, be time consuming and computationally expensive.

Given the underlying complexity in pattern analysis in structured domains, it is safe to say that (for a thoughtful modelling) one needs to consider different strategies for representing and processing data by means of techniques operating at a level closer to the semantics of the data itself. To this aim, we consider processing procedures able to a) automatically find a suitable embedding towards a vector space and b) accommodate the observation above, which rely on the information granulation paradigm and the Granular Computing (GrC) framework [4]–[6]. GrC is an information processing approach that has rapidly expanded in the last decade and has also been successfully employed to synthesise advanced pattern recognition systems suitable for dealing in structured domains such as graphs [7]–[10], sequences [11], [12] and images [13], [14]. One of the most intriguing peculiarities of GrC-based structural pattern recognition systems lies on the model interpretability: in fact, the information granules (i.e., atomic entities endowed with high discriminative power) can be analysed by field-experts to gather further insights on the underlying process. In order to limit the human efforts in analysing information granules, one typically aims at selecting the smallest subset of symbols (granules) that, at the same time, maximise the pattern recognition system performances. Besides its usefulness as knowledge discovery is concerned, this *modus operandi* can be boosted by considering also the structural complexity of the synthesised model, whose investigation is the aim of this paper.

In order to investigate how the structural complexity impacts the overall classification model synthesis, three case studies are considered. For the sake of brevity, we will only consider binary classification problems, yet the analysis can easily be performed on any pattern recognition problem (e.g., multi-class classification, function approximation and regression, clustering), as will be stressed in Section VI. The three problems have been considered for being the core of recent works in GrC-based pattern recognition and regard the analysis of real-world biological systems, namely metabolic networks [8] and protein contact networks [9]. Furthermore, the choice stems on them being quite challenging and their relevance

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

within the biology community. In particular, the metabolic networks experiments have the final goal of discriminating different organisms on the sole basis of their metabolic network wiring and, as further stressed in [15], leaded to the promotion of metabolic pathways as 'universal phenotype'. In fact, whilst genotype (in brief, the DNA genome sequence) is a universal feature in every living organism, the same does not hold for phenotypes: indeed, it is pointless to make a universal phenetic classification based on features such as 'the size of the brain' (which is only present in animals) or 'the shape of leaves' (animals have no leaves). Conversely, metabolism can be found in all living organisms, regardless of their position in the biological organisation and biological taxonomy. The two other experiments regard the discrimination of enzymatic proteins vs. non-enzymatic ones and the discrimination of soluble proteins vs. non-soluble ones starting from their residue contact network, a minimalistic representation of the 3D folded state of a protein [16]. These two problems are very hard to solve because no biochemical predictive theory is currently available. In the first problem, the difficulty resides in the fact that the functional classification of enzymes is inherently sloppy: the chemical reactions defining the classes are largely superimposable and in any case they mainly influence a minor part of the protein structure (active site). The second problem is very hard as well because the solubility of a protein *in vivo* is largely determined by the concurrent presence of other proteins that influence the folding of the target protein, while the specific reference database [17] is based on 'isolated proteins' *in vitro* (intrinsic solubility) and has to do with the folding prediction problem that is still out of reach in protein science [18].

II. Case Studies

A. Eukaryotes vs. Prokaryotes Metabolic Pathways

The first case study, originally investigated in [8], aims at classifying different organisms at different taxonomical scales thanks to their respective metabolic pathways. Amongst the four problems discussed in [8], we focus on the discrimination between metabolic pathways belonging to either eukaryotes or prokaryotes organisms. Metabolic pathways can be conveniently described by networks, where nodes correspond to metabolites (product/substrate of a chemical reaction) and edges exist between any two nodes whether there exist a chemical reaction transforming the two metabolites into one another [8], [19]. The information granulation procedure relies on a unified index called INDVAL which accounts for the specificity and the sensitivity of each single chemical reaction with respect to the problem-related classes. The INDVAL score has been originally proposed in [20] for spotting representative species in different environmental condition and its philosophy is straightforward: a given species s is representative, hence useful for the recognition of a given environmental condition ec if it satisfies both of the following properties

1) s must be present in only (or almost only) in the ec-positive objects
2) s must be present in all (or the great majority of) the ec-positive objects.

We re-adapted this idea in order to spot signature substructures in structured data (i.e., chemical reactions in a metabolic network). To this end, the INDVAL score I can be formally defined as:

$$A_{i,j} = \frac{\text{\# graphs having edge } i \text{ in group } j}{\text{\# graphs having edge } i} \tag{1}$$

$$B_{i,j} = \frac{\text{\# graphs having edge } i \text{ in group } j}{\text{\# graphs in group } j} \tag{2}$$

$$I_{i,j} = A_{i,j} \cdot B_{i,j} \cdot 100 \tag{3}$$

where 'edges' are 'chemical reactions', 'graphs' are 'metabolic networks' and 'groups' can be deduced by the problem-related class labels. By definition, since $A_{i,j} \in [0,1]$ and $B_{i,j} \in [0,1]$, then $I_{i,j} \in [0,100]$. The two supporting scores A and B have a straightforward interpretation: the maximum value of A is obtained when the i^{th} edge can be found only in patterns (graphs) belonging to class j, whereas the maximum value for B is obtained if all patterns of class j have edge i. Finally, the maximum INDVAL I corresponds to the maximum sensitivity and specificity for the i^{th} edge within group j: all patterns of class j have edge i and no patterns belonging to other classes have edge i.

Given these preliminary definitions, the granulation procedure can be summarised as follows: let $\mathcal{D}$ be the dataset, properly split into three disjoint and non-overlapping training, validation and test set ($\mathcal{D}_{\text{TR}}$, $\mathcal{D}_{\text{VL}}$, $\mathcal{D}_{\text{TS}}$). Since classification problems are of interest, let $\mathbf{l}$ be the vector containing the corresponding class label for each of the patterns in $\mathcal{D}$ and consider $\mathbf{l}$ to be split accordingly ($\mathbf{l}_{\text{TR}}$, $\mathbf{l}_{\text{VL}}$, $\mathbf{l}_{\text{TS}}$). Let $\mathcal{E}$ be the set of unique edges in $\mathcal{D}_{\text{TR}} \cup \mathcal{D}_{\text{VL}}$ and let $\mathcal{L}$ be the set of problem-related classes, then one can figure $\mathbf{A}, \mathbf{B}, \mathbf{I} \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{L}|}$ as a compact matrix representation of Eqs. (1)–(3).

The next step is to filter (preserve) edges relevant for characterising the different problem-related classes in order to properly build an embedding space: to this end, the end-user defines a threshold $T \in (0,100)$ and only edges in $\mathcal{E}$ having INDVAL score greater than (or equal to) T are included in the alphabet $\mathcal{A}$. In other words, edges are selected if their corresponding row from $\mathbf{I}$ has scores greater than (or equal to) T for at least one of the problem-related classes (columns). After the filtering procedure (i.e., alphabet definition), the embedding procedure can take place by building the symbolic histograms: that is, each pattern is individually transformed in an $|\mathcal{A}|$-length integer-valued vector which, in position i, counts the number of occurrences of the i^{th} symbol in $\mathcal{A}$ within the pattern itself. This embedding procedure moves the problem from the structured domain towards a Euclidean space or, in other words, the three sets $\mathcal{D}_{\text{TR}}$, $\mathcal{D}_{\text{VL}}$ and $\mathcal{D}_{\text{TS}}$ are individually cast into three instance matrices $\mathbf{D}_{\text{TR}} \in \mathbb{R}^{|\mathcal{D}_{\text{TR}}| \times |\mathcal{A}|}$, $\mathbf{D}_{\text{VL}} \in \mathbb{R}^{|\mathcal{D}_{\text{VL}}| \times |\mathcal{A}|}$ and $\mathbf{D}_{\text{TS}} \in \mathbb{R}^{|\mathcal{D}_{\text{TS}}| \times |\mathcal{A}|}$. In the embedding (Euclidean) space, the classifier synthesis can take place. Synthesising the model should address two important aspects:

1) find a suitable set of hyperparameters $\mathcal{H}$ for the classifier
2) find a suitable subset of meaningful features

in order to maximise the classifier performances. A genetic algorithm can be employed in this regard [21], where the genetic code has the form

$$\begin{bmatrix}\mathcal{H} & \mathbf{w}\end{bmatrix} \tag{4}$$

where $\mathbf{w} \in \{0,1\}^{|\mathcal{A}|}$ is the boolean feature selection vector in charge of discarding unpromising features (i.e., symbols from $\mathcal{A}$). The fitness function driving the genetic optimisation shall therefore take into account both the classifier performances π (on the validation set) and the sparsity σ of the feature selection vector:

$$F = \alpha \cdot \pi + (1-\alpha) \cdot \sigma \tag{5}$$

At the end of the optimisation procedure, the final performances are evaluated on the test set.

As the dataset is concerned, from the KEGG database [22] we dumped the 5299 organisms for which the metabolic network is known and marked each network with a ground-truth class label according to the cellular architecture of the organism itself: 'eukaryote' (439 organisms) or 'prokaryote' (4860 organisms). For building the alphabet, a threshold value of $T = 50$ has been used.

B. Enzymatic vs. Non-Enzymatic Proteins

The second case study tackles another real-world biological system conveniently modelled by a network: proteins. Indeed, the folded state of a protein can be conveniently described by its residue contact network [16], where nodes correspond to amino-acids and edges exist whether they are in spatial proximity (i.e., the Euclidean distance between α-carbon atom locations is within $[4, 8]$Å [3], [9], [23]–[25]). This case study stems from a previous work [9] where the embedding space has been build thanks to the simplicial decomposition of the 3-dimensional contact network with the final goal of predicting whether the protein is an enzyme or not (i.e., it has been assigned an Enzyme Commission (EC) number or not [26]). As for the previous case study, let $\mathcal{D}_{\text{TR}}$, $\mathcal{D}_{\text{VL}}$ and $\mathcal{D}_{\text{TS}}$ be three non-overlapping splits of the dataset $\mathcal{D}$ composed by 3-dimensional networks. Now let each network belonging to each of the three sets be individually decomposed into its clique complex [27]–[29], which encodes the same information as the underlying graph, but additionally completes the skeletal network to its fullest possible simplicial structure. The clique complex is formally defined as the simplicial complex formed by the set of cliques of the underlying graph or, in other words, as the topological space in which each k-vertex clique is represented by a $(k-1)$-simplex. Finally, let each vertex belonging to each simplex forming each simplicial complex to be identified by a categorical attribute: in this case, the amino-acid type. It is straightforward then to evaluate the set of unique simplices belonging to the simplicial complexes of $\mathcal{D}_{\text{TR}} \cup \mathcal{D}_{\text{VL}}$: this set composes the alphabet $\mathcal{A}$. Given $\mathcal{A}$, the embedding procedure may take place: instead of counting edges in a graph (Section II-A), building the symbolic histogram now reads as counting simplices in simplicial complexes.

As per the metabolic networks case, a genetic optimisation can be employed to fully automatise the model synthesis (cf. Eqs. (4)–(5)).

As the dataset is concerned, from UniProt [30] we dumped the entire Escherichia coli str. K12 proteome and cross-checked the list with Protein Data Bank [31] in order to consider only resolved proteins (namely, proteins whose folded structure is known). Then, we performed the following data filtration:

1) proteins with multiple EC numbers have been discarded
2) in PDB files having multiple models, only the first has been considered
3) for atoms having multiple locations, only the first has been considered
4) in order to consider only good quality structures, proteins having PDB files with no information about measurement resolution have been discarded and, similarly, proteins having measurement resolution greater than 3Å have been discarded as well.

A suitable set of 5583 proteins is returned, which have been marked as 'enzymes' (3702 proteins) if they have been provided with an EC numbers or 'not-enzymes' (1181 proteins) otherwise.

C. Soluble vs. Non-Soluble Proteins

This third case study is methodologically equivalent to the one presented in Section II-B and still regards protein networks and their simplicial structure. However, target of the learning system is to classify whether proteins are soluble or not (i.e., they tend to fold by themselves or not). Whilst the EC number is categorical by definition, the solubility degree is a real-valued scalar which, after straightforward normalisation, can be considered spanning the range $[0, 1]$.

The data retrieval process consisted in a cross-check between the eSol database[1] containing the solubility degree (in percentage) for the *E. coli* proteins using the chaperone-free PURE system [32] and the Protein Data Bank in order to retrieve the structure files. All proteins having solubility degree greater than 100% have been thresholded as 100%[2] and, after straightforward normalisation, the solubility degree can be considered as a real-valued scalar in range $[0, 1]$. The four filtering steps as per the EC number case have been performed as well, leading to a total number of 4781 proteins. Finally, the solubility degrees have been thresholded using 0.6 as cutoff value (according to [9]) in order to mark 'soluble' (2421 proteins) vs. 'non-soluble' (2360 proteins).

III. Taking the Model Structural Complexity into Account

In all of the three case studies we considered the following two quality factors for the automatic model synthesis: sparsity

[1] http://tp-esol.genes.nig.ac.jp/ developed in the Targeted Proteins Research Project.

[2] The (small) deviations from 100% can be ascribed to minor experimental errors.

of the feature selection vector and classifier performances. Whilst the latter is straightforward, a major emphasis has been put towards a further refinement (filtering) of the relevant symbols due to the following practical issues:

- the interpretability of the model greatly improves (e.g., less symbols to be analysed by field-experts)
- testing new patterns is faster (i.e., less symbols to match with in order to build its symbolic histogram).

In the reference works (see [8] for metabolic networks and [9] for protein networks), a standard tradeoff value of $\alpha = 0.5$ has been used in order to give the same importance to both quality actors in the fitness function (see Eq. (5)). Undoubtedly, this leaded to a smaller number of surviving information granules (yet, with a minor performance decay) with respect to the common scenario in which one aims at maximising the classifier performances (i.e., $\alpha = 1$). This observation somewhat links to the first practical issue above: the workload for field-experts (biologists, in this case) in analysing the resulting symbols (either be chemical reactions or simplices) was way lower. Nonetheless, there is a third player that somehow goes unnoticed when it comes to design GrC-based classification systems (and still is unexplored in the literature) because one certainly wants to enhance the human-interpretable peculiarity of GrC: the model structural complexity. In fact, there is no a-priori correlation between a low-dimensional space (i.e., few selected symbols) and the smoothness of the decision boundary. In this analysis, we want to consider this third player into account by slightly revisiting the fitness function in the model synthesis phase. In fact, whilst the former (see Eq. (5)) considered only the model performances π and the dimensionality of the embedding space σ, it can be generalised as follows

$$F = \alpha \cdot \pi + \beta \cdot \sigma + \delta \cdot \kappa \tag{6}$$

where κ considers the structural complexity of the model and the triad $\langle \alpha, \beta, \delta \rangle$ is in charge of weighting each term. By assuming that $\pi, \sigma, \kappa \in [0, 1]$, we further let $\alpha, \beta, \delta \in [0, 1]$ in order to ensure a fair contribution amongst π, σ and κ.

IV. A Manual Investigation

A first investigation considers in manually setting the values for α, β and δ in the genetic optimisation phase.

Let us define the three objective in the fitness first. According to previous works [8], [9] the informedness J has been selected as a suitable performance index [33] being one of the very few unbiased indices in case of heavily unbalanced classification problems [34]. The informedness is defined as

$$J = \text{Specificity} + \text{Sensitivity} - 1, \qquad J \in [-1, 1] \tag{7}$$

For consistency with later objectives, a normalised version is adopted

$$\overline{J} = \frac{1}{2}(\text{Specificity} + \text{Sensitivity}), \qquad \overline{J} \in [0, 1] \tag{8}$$

Since we seek to minimise the fitness function, the performance term π reads as:

$$\pi = 1 - \overline{J}, \qquad \pi \in [0, 1] \tag{9}$$

The sparsity term σ is trivially given by the ratio of selected symbols, hence:

$$\sigma = \frac{|\{i : \mathbf{w}_i = 1\}|}{|\mathbf{w}|}, \qquad \sigma \in [0, 1] \tag{10}$$

The structural complexity term κ, alike the set of hyperparameters $\mathcal{H}$ to be optimised by the genetic algorithm (cf. Eq. (4)), is strictly classifier-dependent. For our tests, we chose Support Vector Machines (SVMs) and, specifically, the ν-SVM formulation [35] equipped with a radial basis function kernel of the form[3]

$$K(\mathbf{a}, \mathbf{b}; \mathbf{w}) = \exp\{-\gamma d(\mathbf{a}, \mathbf{b}; \mathbf{w})^2\} \tag{11}$$

where $d(\mathbf{a}, \mathbf{b}; \mathbf{w})$ reads as the weighted Euclidean distance between the two generic patterns $\mathbf{a}$ and $\mathbf{b}$, with $\mathbf{w}$ acting as weighting vector. Formally,

$$d(\mathbf{a}, \mathbf{b}; \mathbf{w}) = \sqrt{\sum\nolimits_{i=1}^{n} \mathbf{w}_i (\mathbf{a}_i - \mathbf{b}_i)^2} \tag{12}$$

being n the size of the considered vectors (i.e., $n = |\mathcal{A}|$). For SVM, the number of support vectors (SVs) computed during the training phase indicates its structural complexity and is also strictly related to the testing phase computational complexity. The latter is indeed linear with the number of SVs, since the SVM decision for a previously-unseen pattern $\mathbf{x}$ is computed as the sign of

$$f(\mathbf{x}) = \sum\nolimits_{i=1}^{\#\text{SVs}} \mu_i y_i K(\mathbf{s}^{(i)}, \mathbf{x}) + b \tag{13}$$

where $\mathbf{s}^{(i)}$ is the i^{th} SV, y_i and μ_i are its class label and the Lagrange multiplier associated to it and b depicts the intercept of the separating hyperplane. In conclusion, the structural complexity term κ reads as:

$$\kappa = \frac{\text{number of SVs}}{|\mathcal{D}_{\text{TR}}|}, \qquad \kappa \in [0, 1]. \tag{14}$$

Given the definitions of π, σ and κ, for each of the three case studies, three weighting setups have been considered:

- $\alpha = \beta = \delta = 1/3$, in order to give the same importance to the three targets
- $\alpha = \beta = 1/2$ and $\delta = 0$, in order to neglect the structural complexity
- $\alpha = \delta = 1/2$ and $\beta = 0$, in order to neglect the sparsity.

Obviously the possibility of having $\alpha = 0$ has not been considered, as this would neglect the classifier performances. The genetic algorithm has been configured to host 100 individuals per 100 generations with a strict early-stop criterion if the average fitness function over $1/3^{\text{rd}}$ of the total number of generations is less than or equal to 10^{-6}, the elitism is set to 10% of the population, the selection follows the roulette wheel heuristic, the crossover operator generates new offsprings in a scattered fashion and the mutation acts in a flip-the-bit fashion for boolean genes ($\mathbf{w}$) and adds to real-valued genes ($\mathcal{H} = \{\nu, \gamma\}$) a random number extracted from a zero-mean

[3]Despite the weighting vector $\mathbf{w}$, the kernel satisfies Mercer's condition: since $\mathbf{w}$ is a boolean vector, the proof is trivial.

TABLE I
STARTING EMBEDDING SPACE CARDINALITY

Case Study	Alphabet Size (avg)	Alphabet Size (std)
Metabolic Pathways	622.8	6.9785
Protein Contact Networks (EC number)	12012.2	33.8112
Protein Contact Networks (Solubility)	11576.6	33.8718

TABLE II
CASE STUDY 1: METABOLIC PATHWAYS

Weights	Informedness	Sparsity	Complexity
$\alpha = 1/3$, $\beta = 1/3$, $\delta = 1/3$	0.9987 ± 0.0020	29.9073 ± 5.4064	0.2717 ± 0.0940
$\alpha = 1/2$, $\beta = 1/2$, $\delta = 0$	0.9973 ± 0.0041	24.3153 ± 3.9548	4.6113 ± 2.4806
$\alpha = 1/2$, $\beta = 0$, $\delta = 1/2$	0.9981 ± 0.0040	49.6321 ± 2.9112	0.1660 ± 0.0338

TABLE III
CASE STUDY 2: PROTEIN CONTACT NETWORKS (EC NUMBER)

Weights	Informedness	Sparsity	Complexity
$\alpha = 1/3$, $\beta = 1/3$, $\delta = 1/3$	0.7704 ± 0.0346	9.7153 ± 4.0119	16.7037 ± 4.4044
$\alpha = 1/2$, $\beta = 1/2$, $\delta = 0$	0.8588 ± 0.0185	8.0401 ± 1.5319	44.6578 ± 3.4921
$\alpha = 1/2$, $\beta = 0$, $\delta = 1/2$	0.7751 ± 0.0344	24.7765 ± 20.4177	17.1408 ± 4.1058

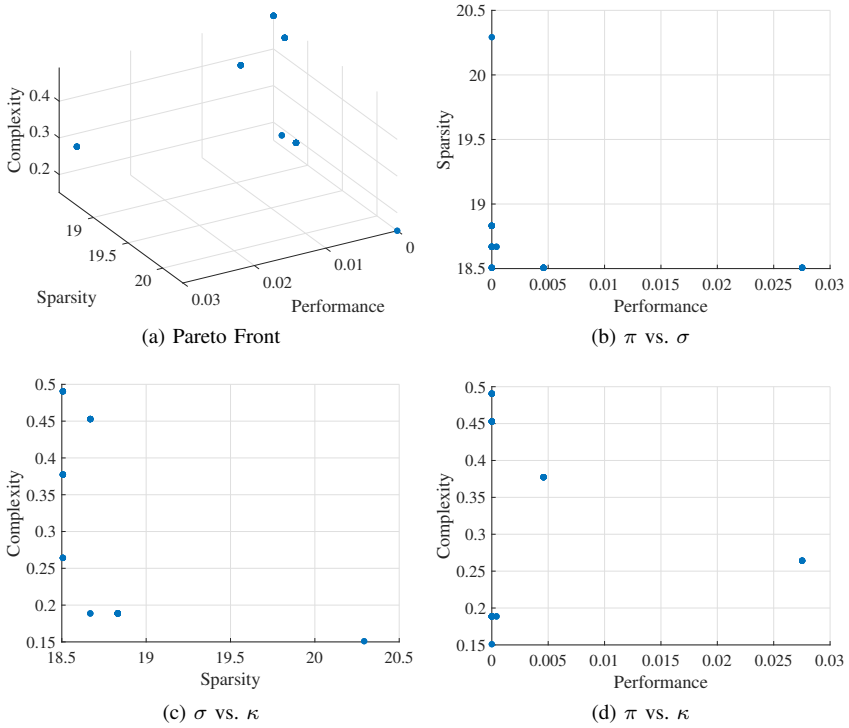
TABLE IV
CASE STUDY 3: PROTEIN CONTACT NETWORKS (SOLUBILITY)

Weights	Informedness	Sparsity	Complexity
$\alpha = 1/3$, $\beta = 1/3$, $\delta = 1/3$	0.8361 ± 0.0120	8.7649 ± 3.0631	16.0586 ± 1.9406
$\alpha = 1/2$, $\beta = 1/2$, $\delta = 0$	0.9082 ± 0.0144	6.9236 ± 1.0333	46.2008 ± 6.1070
$\alpha = 1/2$, $\beta = 0$, $\delta = 1/2$	0.8376 ± 0.0303	29.1630 ± 12.4602	14.3013 ± 2.0193

Gaussian distribution whose variance shrinks as generations go by. For the three case studies, five stratified training-validation-test splits have been performed and, in order to ensure a comparison as fair as possible, the same splits have been fed for each weights setup. Table I shows the alphabet size, in terms of number of symbols (average and standard deviation across the five splits). Tables II, III and IV show the results on the test set for the three case studies from Section II. Results are presented in terms of average and standard deviation across the five training-validation-test splits and regard informedness, sparsity (i.e., percentage of selected symbols) and complexity (i.e., percentage of patterns elected as SVs). Case Study 1 (Table II) seems to be the easiest to solve amongst the three: regardless of α, β and δ, performances are always greater than 99% and, when sparsity is weighted ($\beta \neq 0$), approximately 24-25% of the symbols are selected, against a 49% of symbols when sparsity is neglected. The shift for structural complexity is much lower: less than 1% of the training patterns are elected as SVs when $\delta \neq 0$ against a 4% when complexity is neglected. Case Studies 2 and 3 are much harder to solve. There is a non-negligible performance drop when $\delta \neq 0$ (approx. 7%); however, when $\beta \neq 0$ there is a clear-cut improvement (approx. 20%) in terms of sparsity and when $\delta \neq 0$ the complexity improves of approx. 30%. Obviously, there is no clear winner here and the choice of $\langle \alpha, \beta, \delta \rangle$ is strictly scenario- and application- related.

V. AN AUTOMATIC INVESTIGATION

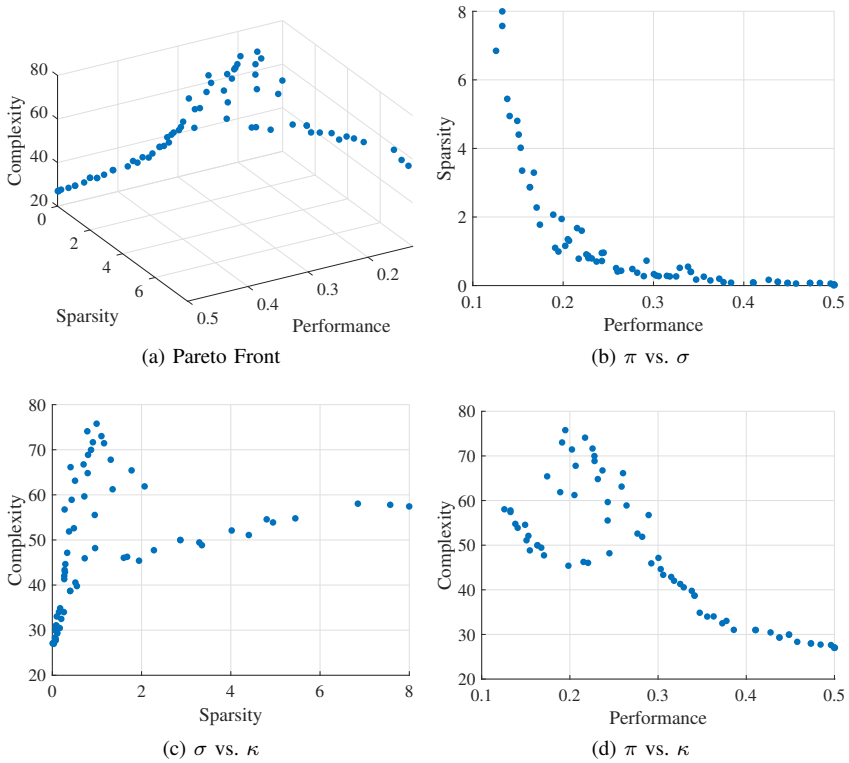
The 'no-winner dilemma' of the previous Section highlights the peculiar non-dominant multiobjective nature of the problem at hand. Given this viewpoint, a second investigation relies on tracking the Pareto front in the 3D space spanned by the three objective functions π, σ and κ. To this end, a controlled-elitist genetic algorithm [36] has been employed. Conversely to standard elitist genetic algorithms, which aim at promoting individuals with better fitness function, controlled-elitist genetic algorithms favour individuals which help in promoting the diversity of the population. The multi-objective genetic algorithm has been configured to host 200 individuals in each generation with a maximum number of $200 \cdot (|\mathcal{H}| + |\mathcal{A}|)$ generations. The early-stop criterion must consider the behaviour of the population on the Pareto front: the algorithm stops when the geometric average of the relative change in value of the spread (i.e., the movement of the Pareto set) over 100 generations is less than 0.0001, and the final spread is less than the mean spread over the past 100 generations. Fig. 1–3 show the Pareto front and the pairwise 2D projections for the three case studies, respectively. Case Study 1 shows a very compact Pareto front with very few surviving individuals very close to the origin (recall that the three objective functions



(a) Pareto Front (b) π vs. σ

(c) σ vs. κ (d) π vs. κ

Fig. 1. Case Study 1: Metabolic Pathways.

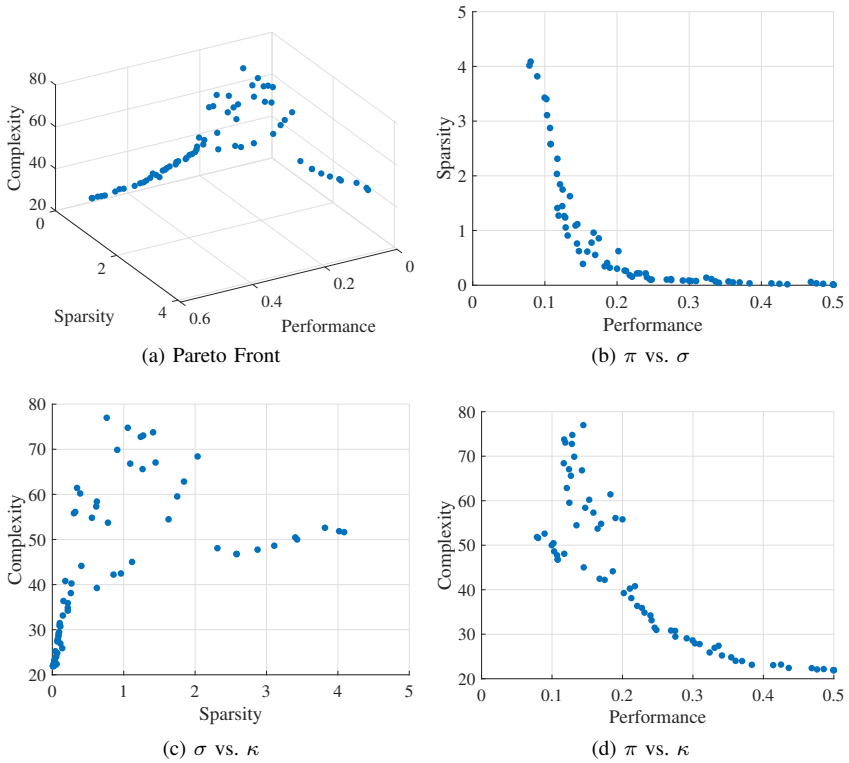


(a) Pareto Front (b) π vs. σ

(c) σ vs. κ (d) π vs. κ

Fig. 2. Case Study 2: Protein Contact Networks (EC number).

must be minimised). Conversely, the Pareto front for Case Studies 2 and 3 (Fig. 2a–3a) show a peculiar front with two very smooth sections and a noisy section in-between: for Case Study 3, the noisy part is observed for $\pi \in [0.23, 0.13]$, whereas for Case Study 2 we have $\pi \in [0.3, 0.17]$. For all case studies, the sparsity-vs-performances projection (Fig. 1b, 2b, 3b) shows a very neat elbow-shaped front with straightforward interpretation: performance degrades if too few features



(a) Pareto Front

(b) π vs. σ

(c) σ vs. κ

(d) π vs. κ

Fig. 3. Case Study 3: Protein Contact Networks (Solubility).

are selected. Indeed, very few features might not properly characterise the decision boundary, with the extreme case where $\sigma \rightarrow 0$, yielding $\pi \rightarrow 0.5$, namely the behaviour of a random classifier ($\pi = J = 0$ in the un-normalised case). If we leave the noisy blob aside, Fig. 2d and 3d show a somewhat linearly decreasing behaviour when it comes to relate complexity and performance, whereas Fig. 2c and 3c show a somehow increasing trend (complexity-vs-sparsity). Nonetheless, the part where the noisy blob appears is indeed the interesting part of the 3D Pareto front, being the part where the three objective functions tend to have overall lower values. Specifically, for Case Study 2 the noisy part of the front lies in $\pi \times \sigma \times \kappa \simeq [0.15, 0.25] \times [0.2, 2] \times [45, 70]$, whereas for Case Study 3 the noisy part of the front lies in $\pi \times \sigma \times \kappa \simeq [0.1, 0.2] \times [0.3, 1.5] \times [40, 75]$.

VI. Conclusions

In this paper, we focused on embedding spaces optimisation for structured pattern recognition, following a GrC approach based on symbolic histograms. Alongside the well-known joint optimisation of performances and number of relevant information granules, a third player has been included: the structural complexity of the trained model. Such a measure, in fact, is closely related to the smoothness of decision surface synthesised during training. As suggested by regularisation theory, minimising the structural complexity of the model can avoid the overfitting phenomenon, boosting the generalisation capability of the final classification model. Clearly, minimising at the same time the embedding space dimensionality, the structural complexity of the model and the performance (in our case, defined as the complement of informedness) means facing a proper multi-objective optimisation problem, since they are conflicting functions. In order to study the relationships between these three objective functions, three graph-based problems have been considered for their biological significance, their increasing levels of difficulty and for representing different ways of synthesising pivotal information granules. In a first step of our analysis, we have defined a weighted convex linear combination as the overall fitness, fixing in advance the relative importance of each objective function in the optimisation procedure, showing that introducing complexity allows better performances on test set, at the expense of an increase of the embedding dimension (Case Study 1). However, when classes are much more overlapped a different behaviour can be observed (Case Studies 2 and 3). For a deeper analysis, we have adopted a tailored stochastic multi-objective optimisation procedure in order to evolve solutions towards the definition of a Pareto front. This automated analysis confirms that Case Studies 2 and 3 are much more challenging than the first one, showing that in these cases solutions close to the origin in the three-dimensional objective function space does not show a clear trend, spreading chaotically in a cluster of best candidates. This is a clear hint that in such difficult classification problems the existence of multiple solutions for the optimal embedding musk and the wide superposition of decision regions can give rise to clusters of solutions close to the origin characterised by large entropy values, due to strong non-linear correlation

effects between conflicting objective functions. As concerns future works, in this paper, for the sake of simplicity, we have faced only binary classification problems, yet extension towards multi-class problems is straightforward. Even more, π can be personalised also in order to accommodate other machine learning problems, as clustering (considering relative validation indices such as Davies-Bouldin or the Silhouette [37]) or function approximation problems (mean squared error, coefficient of determination), following a similar optimisation approach based on information granulation. Moreover, we plan to investigate suitable entropy-based measures to study possible correlations between the chaotic spreading of best candidate solutions (closed to the origin) and the degree of difficulty of the classification problem at hand, in terms of both the embedding mapping complexity and roughness of decision surfaces.

REFERENCES

[1] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *SIGKDD Explor. Newsl.*, vol. 12, no. 1, p. 40–48, 2010.

[2] A. Martino, A. Giuliani, and A. Rizzi, "Granular computing techniques for bioinformatics pattern recognition problems in non-metric spaces," in *Computational Intelligence for Pattern Recognition*, ser. Studies in Computational Intelligence, W. Pedrycz and S.-M. Chen, Eds. Cham: Springer International Publishing, 2018, no. 777, pp. 53–81.

[3] A. Martino, A. Rizzi, and F. M. Frattale Mascioli, "Supervised approaches for protein function prediction by topological data analysis," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[4] A. Bargiela and W. Pedrycz, "Toward a theory of granular computing for human-centered information processing," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 2, pp. 320–330, 2008.

[5] ——, "The roots of granular computing," in *2006 IEEE International Conference on Granular Computing*, 2006, pp. 806–809.

[6] ——, *Granular computing: an introduction*. Kluwer Academic Publishers, Boston, 2003.

[7] L. Baldini, A. Martino, and A. Rizzi, "Towards a class-aware information granulation for graph embedding and classification," in *Computational Intelligence: 11th International Joint Conference, IJCCI 2019 Vienna, Austria, September 17-19, 2019 Revised Selected Papers*, To appear in.

[8] A. Martino, A. Giuliani, V. Todde, M. Bizzarri, and A. Rizzi, "Metabolic networks classification and knowledge discovery by information granulation," *Computational Biology and Chemistry*, vol. 84, p. 107187, 2020.

[9] A. Martino, A. Giuliani, and A. Rizzi, "(hyper)graph embedding and classification via simplicial complexes," *Algorithms*, vol. 12, no. 11, 2019.

[10] L. Baldini, A. Martino, and A. Rizzi, "Stochastic information granules extraction for graph embedding and classification," in *Proceedings of the 11th International Joint Conference on Computational Intelligence - Volume 1: NCTA, (IJCCI 2019)*, INSTICC. SciTePress, 2019, pp. 391–402.

[11] E. Maiorino, F. Possemato, V. Modugno, and A. Rizzi, "Information granules filtering for inexact sequential pattern mining by evolutionary computation," in *Proceedings of the International Joint Conference on Computational Intelligence - Volume 1*, ser. IJCCI 2014. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda, 2014, p. 104–111.

[12] ——, "Noise sensitivity of an information granules filtering procedure by genetic optimization for inexact sequential pattern mining," in *Computational Intelligence*, J. J. Merelo, A. Rosa, J. M. Cadenas, A. Dourado, K. Madani, and J. Filipe, Eds. Cham: Springer International Publishing, 2016, pp. 131–150.

[13] A. Rizzi and G. Del Vescovo, "Automatic image classification by a granular computing approach," in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006, pp. 33–38.

[14] G. Del Vescovo and A. Rizzi, "Online handwriting recognition by the symbolic histograms approach," in *2007 IEEE International Conference on Granular Computing (GRC 2007)*, 2007, pp. 686–686.

[15] A. Martino, A. Giuliani, and A. Rizzi, "The universal phenotype," *Organisms. Journal of Biological Sciences*, vol. 3, no. 2, 2019.

[16] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani, "Protein contact networks: an emerging paradigm in chemistry," *Chemical Reviews*, vol. 113, no. 3, pp. 1598–1613, 2012.

[17] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi, "Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins," *Proceedings of the National Academy of Sciences*, vol. 106, no. 11, pp. 4201–4206, 2009.

[18] S. H. P. de Oliveira and C. M. Deane, "Exploring folding features in protein structure prediction," *Biophysical Journal*, vol. 114, no. 3, Supplement 1, p. 36a, 2018.

[19] K. Tun, P. K. Dhar, M. C. Palumbo, and A. Giuliani, "Metabolic pathways variability and sequence/networks comparisons," *BMC bioinformatics*, vol. 7, no. 1, p. 24, 2006.

[20] M. Dufrêne and P. Legendre, "Species assemblages and indicator species: the need for a flexible asymmetrical approach," *Ecological monographs*, vol. 67, no. 3, pp. 345–366, 1997.

[21] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[22] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[23] E. Maiorino, A. Rizzi, A. Sadeghian, and A. Giuliani, "Spectral reconstruction of protein contact networks," *Physica A: Statistical Mechanics and its Applications*, vol. 471, pp. 804–817, 2017.

[24] E. De Santis, A. Martino, A. Rizzi, and F. M. Frattale Mascioli, "Dissimilarity space representations and automatic feature selection for protein function prediction," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[25] A. Martino, E. Maiorino, A. Giuliani, M. Giampieri, and A. Rizzi, "Supervised approaches for function prediction of proteins contact networks from topological structure information," in *Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I*, P. Sharma and F. M. Bianchi, Eds. Cham: Springer International Publishing, 2017, pp. 285–296.

[26] E. C. Webb, *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.*, 6th ed. Academic Press, 1992.

[27] D. Horak, S. Maletić, and M. Rajković, "Persistent homology of complex networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 03, p. P03034, 2009.

[28] H.-J. Bandelt and V. Chepoi, "Metric graph theory and geometry: a survey," *Contemporary Mathematics*, vol. 453, pp. 49–86, 2008.

[29] C. Giusti, R. Ghrist, and D. S. Bassett, "Two's company, three (or more) is a simplex," *Journal of Computational Neuroscience*, vol. 41, no. 1, pp. 1–14, 2016.

[30] The UniProt Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, 2017.

[31] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[32] Y. Shimizu, A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa, and T. Ueda, "Cell-free translation reconstituted with purified components," *Nature biotechnology*, vol. 19, no. 8, p. 751, 2001.

[33] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[34] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[35] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[36] K. Deb, *Multi-objective optimization using evolutionary algorithms*. Chichester, England: John Wiley & Sons, 2001.

[37] A. Martino, A. Rizzi, and F. M. Frattale Mascioli, "Distance matrix precaching and distributed computation of internal validation indices in k-medoids clustering," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.