A Stable Hybrid Method for Feature Subset Selection using Particle Swarm Optimization with Local Search

Hassen Dhrif
Department of Computer Science
University of Miami
Miami, Florida
hassen.dhrif@miami.edu

Luis G. S. Giraldo
Department of Computer Science
University of Miami
Miami, Florida
lgsanchez@cs.miami.edu

Miroslav Kubat
Electrical and Computer Engineering Department
University of Miami
Miami, Florida
mkubat@miami.edu

Stefan Wuchty
Department of Computer Science
University of Miami
Miami, Florida
wuchtys@cs.miami.edu

ABSTRACT

The determination of a small set of biomarkers to make a diagnostic call can be formulated as a feature subset selection (FSS) problem to find a small set of genes with high relevance for the underlying classification task and low mutual redundancy. However, repeated application of a heuristic, evolutionary FSS technique usually fails to produce consistent results. Here, we introduce COMB-PSO-LS, a novel hybrid (wrapper-filter) FSS algorithm based on Particle Swarm Optimization (PSO) that features a local search strategy to select the least dependent and most relevant feature subsets. In particular, we employ a Randomized Dependence Coefficient (RDC)-based filter technique to guide the search process of the particle swarm, allowing the selection of highly relevant and consistent features. Classifying cancer samples through patient gene expression profiles, we found that COMB-PSO-LS provides highly stable and non-redundant gene subsets that are relevant for the classification process, outperforming standard PSO methods.

CCS CONCEPTS

• **Theory of computation → Evolutionary algorithms**; **Bio-inspired optimization**; • **Computing methodologies → Feature selection**;

KEYWORDS

feature subset selection, wrapper, filter, hybrid methods, particle swarm optimization, evolutionary computation, multi-objective optimization, local search, randomized dependence coefficient

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6111-8/19/07…$15.00
https://doi.org/10.1145/3321707.3321816

ACM Reference Format:
Hassen Dhrif, Luis G. S. Giraldo, Miroslav Kubat, and Stefan Wuchty. 2019. A Stable Hybrid Method for Feature Subset Selection using Particle Swarm Optimization with Local Search. In *Genetic and Evolutionary Computation Conference (GECCO '19), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3321707.3321816

1 INTRODUCTION

Expression profiles of genes are correlated in non-linear ways, pointing to the existence of irrelevant and redundant genes for the determination of biomarkers. The problem of finding biomarkers that allow the classification of disease and non-disease cases can be formulated as a feature subset selection problem (FSS). As the dimensionality of gene expression datasets increases, a variety of heuristic methods with relatively low computational complexity have been adopted. In particular, evolutionary algorithms in general and particle swarm optimization (PSO) in particular, are computational efficient methods to tackle such high dimensional search problems, but are prone to the selection of similar and irrelevant features in the final feature subset. Since current PSO-based approaches usually do not consider associations between features to guide the search process, similar and correlated features have a high probability to be selected, limiting the classifier performance. Due to their nondeterministic nature, such heuristic search methods draw different sets of features for the same problem in each run [1, 24, 25]. Although different criteria to assess the stability of results have been applied [28, 31, 35, 42], simple combinations of highly frequent features usually fail to provide better classification results [12, 60]. To tackle such problems, new filter-based multi-variate methods have been introduced in recent years [20, 59, 61]. Largely based on mutual information to evaluate relevance and redundancy of selected features, such filter methods have low computational cost but suffer from low classification accuracy. In contrast, wrapper approaches provide superior classification accuracy, but engender high computational cost when applied to large datasets. As hybrid models combine the advantages of filter and wrapper methods, we introduce a novel hybrid (wrapper-filter) PSO approach by integrating a new local search filter operation, allowing us to fine-tune the search process in an organized fashion. Our method, COMB-PSO-LS (COMBinatorial PSO with Local Search) - an extension

of our COMB-PSO algorithm [10] - utilizes association information to efficiently guide the search process. Our approach allows us to find stable subsets of pertinent features (genes) by eliminating irrelevant features with little or no predictive information and redundant genes that are strongly correlated. Furthermore, our reduction procedure accelerates the learning process, leads to a simple and understandable predictor model, and avoids overfitting [38]. We evaluated the performance of COMB-PSO-LS on three synthetic and three well known cancer specific gene expression datasets. Our results demonstrate that our method improves classification accuracy, reduces the size and improves the stability of the selected features.

2 BACKGROUND AND RELATED WORK

2.1 Feature relevance and redundancy

Given that features are defined as *strongly relevant*, *weakly relevant*, and *irrelevant* [27, 67] for the underlying classification problem, an optimal subset must include all strongly relevant features, may include some weakly relevant ones, but no irrelevant features. Searching for an optimal subset based on the definitions of feature relevance and redundancy is combinatorial in nature. Moreover, the optimal subset can only be defined based on the knowledge of the true data distribution. Since such data characteristics are usually unknown, estimation methods are applied to evaluate relevant and redundant features. To increase the efficiency of FSS methods and overcome their high computational cost, a large and growing body of literature has considered multivariate filter methods [15, 32, 39, 40, 43, 54, 55]. More recent attention has focused on hybrid methods [26, 57], mostly integrating the mutual information-based filter model in the framework of PSO-based wrapper methods. Many non-linear statistical dependence measures have been developed recently [3, 5, 16, 17, 22, 34, 44, 45, 52, 53] to assess feature relevance and dependency. Here, we investigate the effectiveness of *randomized dependence coefficients* as a measure of association information to guide the local search.

2.2 FSS stability

Heuristic evolutionary computation (EC) algorithms tend to select different feature subsets with equal prediction accuracy even when applied to the same data multiple times. Furthermore, even small data perturbations such as removal or addition of new data may further prompt algorithms to find different subsets [13, 23, 28, 51, 64]. As a consequence, interpretability of selected subsets is substantially impaired by poor stability. Stability (or robustness) indicates the ability of an algorithm to determine stable feature subset when new training samples are added or removed. In fact, stability of FSS algorithms has received increasing attention [24, 28], indicating that many well-established FSS algorithms suffer from low stability of feature subsets in the presence of small data perturbations.

2.3 Randomized dependence coefficient

The Randomized Dependence Coefficient (RDC) introduced by Lopez-Paz et al. [34], is an empirical estimator of the Hirschfeld-Gebelein-Rényi (HGR) maximum correlation coefficient that measures non-linear dependencies between random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. As an example related to the FSS problem, $p = 1$ or 2 when X represents individual or joint feature vectors respectively, and $q = 1$ when Y corresponds to a class labels represented as scalars. RDC uses two paired sets of m samples, $\mathbf{X} \in \mathbb{R}^{m\times p}$ and $\mathbf{Y} \in \mathbb{R}^{m\times q}$, to approximate the HGR measure (Fig. 1). First, an empirical copula transformation, where the marginal distributions of the resulting variables become uniform, makes the RDC invariant to any strictly monotonic functions applied to any of the dimensions of the input variables. The copula transformation is followed by non-linearly projecting the samples to random sets of functions in $\mathbb{R}^p$ and $\mathbb{R}^q$. The resulting transformed samples are denoted by $\mathbf{\Phi}_X \in \mathbb{R}^{m\times k}$ and $\mathbf{\Psi}_Y \in \mathbb{R}^{m\times \ell}$, where k and ℓ are the dimensions of the projections for X and Y. Finally, the RDC score is defined as:

$$\mathrm{RDC}(X, Y) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^k, \boldsymbol{\beta}\in\mathbb{R}^\ell} \overline{\rho}\left(\boldsymbol{\alpha}^T\mathbf{\Phi}_X, \boldsymbol{\beta}^T\mathbf{\Psi}_Y\right), \tag{1}$$

where $\overline{\rho}$ is the Pearson's correlation coefficient that can be obtained by linearly combining the feature dimensions (rows) of each transformed sample $\mathbf{\Phi}_X$ and $\mathbf{\Psi}_Y$. The rationale behind the two-stage transformation is to ease the hyper-parameter selection of the randomized projections. In particular, we use sine and cosine projections, $(\sin(W^TX + b),\ \cos(W^TX + b))$, as suggested in [34]. In this expansion, W is a zero-mean multivariate Gaussian with covariance $s\boldsymbol{I}$, while b is uniform in $[-\pi, \pi]$. By operating on the copula instead of the original input variables, the choice of the scale parameter s is not influenced by the spread or the position of the original variables. In practice, the number of random projections, k and ℓ, still need some tuning. If the number of dimensions get closer to the number of samples m, the RDC may uncover spurious correlations. In turn, a very small number of dimensions may over-regularize the dependence measure, hindering its ability to capture non-linear dependencies.

2.4 The PSO algorithm

Particle swarm optimization (PSO) algorithm [14], its binary variant (BPSO) [8, 9, 29, 30, 58, 69] and its multi-objective variant MOPSO [6, 37, 62, 66, 68] are evolutionary computation techniques, that have been combined with different classification methods to select informative markers from gene expression data [7, 18, 19, 36, 47, 63]. In particular, the PSO algorithm is based on the concept of moving particles in a search space. At each iteration, a particle's velocity is updated by

$$\vec{v}_i(t+1) = \omega\vec{v}_i(t) + r_1c_1(\vec{p}_i - \vec{x}_i(t)) + r_2c_2(\vec{g} - \vec{x}_i(t)) \tag{2}$$

while a particles location is calculated by

$$\vec{x}_i(t+1) = \vec{x}_i(t) + \vec{v}_i(t) \tag{3}$$

where $\vec{v}_i(t)$ and $\vec{x}_i(t)$ are the velocity and the position of particle i in a n-dimensional search space at the t^{th} iteration, respectively. Vectors $\vec{p}_i$ and $\vec{g}$ are the particle's personal best (*pbest*) and the swarms best (*gbest*) values, respectively. Coefficients c_1 and c_2 are acceleration constants, and r_1 and r_2 are random values. Parameter ω is the inertia weight to control the impact of the last velocity on the current velocity. Usually, velocity values are limited to the range $[-v_{max}, v_{max}]$, by a predefined maximum velocity, v_{max}.

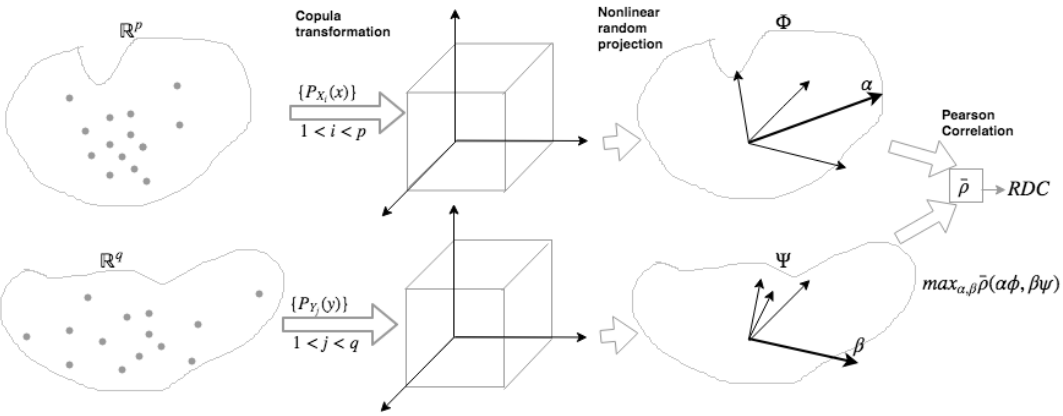

Figure 1: Overview of RDC computation. First, data samples from X and Y are mapped to unit hypercubes in $\mathbb{R}^p$ and $\mathbb{R}^q$ by making marginal distributions of each of their dimensions uniform (copula transformation). The mapped samples are subjected to non-linear maps using random sets of functions Φ and Ψ. Canonical correlation between the mapped samples $\mathbf{\Phi}_X \in \mathbb{R}^{m\times k}$ and $\mathbf{\Psi}_Y \in \mathbb{R}^{m\times \ell}$ is computed to provide the dependence measure.

2.5 The COMB-PSO algorihm

Recently, we introduced a combinatorial variant of PSO, COMB-PSO [10] to analyze gene expression datasets with tens of thousands of genes and few hundreds samples. In the following, we highlight characteristics of COMB-PSO allowing to (i) maximize the accuracy of sample classification, (ii) minimize the underlying set size of selected features, and (iii) maintain stability of the size of feature subsets in the massive presence of uninformative (*i.e.* irrelevant) features. However, COMB-PSO ignores the relevance and redundancy of selected features, limiting the stability of its outcomes.

2.5.1 Improved exploration and exploitation capabilities. In a binary search space, a particle moves by flipping its bits. Such a definition of movement does not provide a very intuitive notion of velocity, making the notions of speed, direction, and momentum to the binary domain less clear. Furthermore, the non-monotonic shape of the changing probability function has a negative effect on the exploitation capability of the algorithm (*i.e.* the use of information gathered from past iterations), and BPSO tend to show poor scaling behavior. To retain the advantage that continuous PSO has superior search capabilities compared to BPSO, COMB-PSO introduces a new binary vector $\vec{b}$ to map the continuous space position to binary digits by

$$b_{ij} = \begin{cases} 1, & \text{if } rand() < S(x_{ij}) \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where

$$S(x_{ij}) = \frac{1}{(1 + e^{-x_{ij}})}, \tag{5}$$

indicating that feature j in particle i is accounted for in a feature subset if $b_{ij} = 1$.

2.5.2 Better transition from exploration to exploitation. To allow a faster transition from exploration to exploitation, COMB-PSO uses a sigmoid function for the inertia weight ω and acceleration coefficients c_1 and c_2, extending the particles time to explore and exploit the search space by

$$\begin{aligned} \omega &= \omega_{min} + (\omega_{max} - \omega_{min})\frac{1}{1 + (\frac{t}{aT})^b} \\ c_1 &= c_{min} + (c_{max} - c_{min})\frac{1}{1 + (\frac{t}{aT})^b} \\ c_2 &= c_{max} + (c_{min} - c_{max})\frac{1}{1 + (\frac{t}{aT})^b}. \end{aligned} \tag{6}$$

Such a function is shown in Fig. 2, where a and b are the transition coefficients. a governs the transition point and b determines the length of the exploration and exploitation phase of the particles. Compared to a linearly decreasing function, the proposed function secures that particles transition fast between full exploration and exploitation modes.

2.5.3 Convergence rate. COMB-PSO introduces a dynamic population strategy where updated best positions (personal and global) are not discarded, but eventually swapped with the weakest performing particles in the decision space.

2.5.4 Avoiding premature convergence trap. Furthermore, COMB-PSO introduces the turbulence coefficients θ and γ. $\gamma \in [0, 1]$ indicates the fraction of particles randomly selected that reset their velocities when *gbest* stagnates after θ consecutive iterations.

2.5.5 Reducing feature subset size. The choice of the boundary values of v_{min} and v_{max} requires some care since they not only affect the balance between exploration and exploitation, but also the size of the generated subsets. If v_{max} is too large, many irrelevant features will be selected. In turn, some critical features will be missed in the selection process if v_{min} is too small. While most methods adopt symmetric boundaries as velocity constraint (*i.e.* $[-v_{max}, v_{max}]$), COMB-PSO introduces the asymmetric coefficient λ as defined in Eq. 7.

$$v_{max} = -\lambda v_{min}, \lambda \in [0, 1], \tag{7}$$

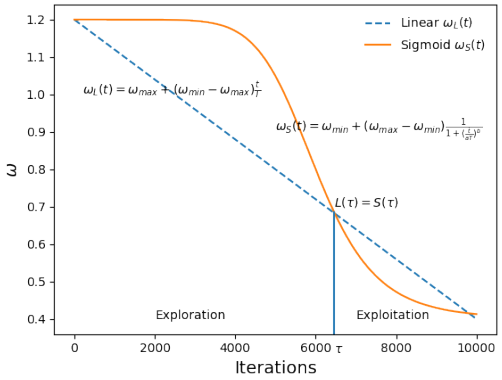

Figure 2: Transition between exploration and exploitation phases. COMB-PSO-LS applies a sigmoid function to establish inertia weights where, ω_{min}=0.4, ω_{max}=1.2, a=0.6, b=8 and T=10 000. We define the break point τ between exploration and exploitation phase as the intersection between a linear and sigmoid function, assuring $0 < \tau \ll T$.

As a consequence, an elevated value of λ increases the probability obtaining additional features.

3 THE METHOD

Although COMB-PSO provides a global search strategy to find near optimal feature subsets, fine-tuning the search near local optima remains a weakness. Here, we introduce a novel hybrid wrapper-filter FSS, COMB-PSO-LS, that combines our previous COMB-PSO search approach with a local search method that is based on determining randomized dependence coefficients (RDC) between features and class labels.

3.1 Association approximation measure

Yu and Liu [67] distinguished between individual and combined associations. Given a dataset of m samples and n features, the individual association between any feature $F_i \in \mathbb{R}^{m\times n}$ and the class $C \in \mathbb{R}^{m\times q}$ is defined by

$$c_i = RDC(F_i, C) \tag{8}$$

where vector $c_i, i < n$, represents the relevance between candidate input features and target output. The combined association between any pair of features F_i and F_j $(i \neq j)$ and the class C is defined by

$$\mathcal{Q}_{ij} = \begin{cases} 0 & i = j \\ RDC(\{F_i, F_j\}, C) & i \neq j, \end{cases} \tag{9}$$

where the square matrix $\mathcal{Q} \in \mathbb{R}^{n\times n}$ indicates the relevance between candidate input features and target output as a measure of redundancy. In our approximation method, we first measure the individual association of each feature and heuristically treat all features as relevant but subjected to redundancy analysis. Approximately determining redundancy between two features as a function of both their individual and combined associations, we assume that a feature with a larger individual association value holds more information about the class than a feature with a smaller individual association value. For two features F_i and F_j with $c_i \geq c_j$, we evaluate whether feature F_i can form an approximate redundant cover for feature F_j (instead of F_j for F_i) to maintain more information about the class. In addition, if combining F_j with F_i does not provide more predictive power in determining the class than using F_i alone, F_i forms an approximate redundant cover for F_j. Such considerations are covered by

Definition 3.1. **(Approximate redundant cover)** For two features F_i and F_j, F_i forms an approximate redundant cover for F_j iff $c_i \geq c_j$ and $c_i \geq Q_{i,j}$.

and

Definition 3.2. **(Predominant feature)** A feature is predominant if it does not have any approximate redundant cover in the current set.

As a consequence, we remove every feature with an approximate redundant cover through a predominant feature through

Predominant features will not be removed at any stage. If a feature F_j is removed based on a predominant feature F_i in an earlier phase, it is guaranteed that it will still find an approximate redundant cover (the same F_i) in any later phase when another feature is removed. Since the feature with the highest individual association value does not have any approximate redundant cover, it must be one of the predominant features and can be used as the starting point to determine the redundancy between the remaining features. Concluding, our approximation method of relevance and redundancy analysis allows us to find all predominant features, while other remaining features are eliminated.

3.2 The local search extension to COMB-PSO, COMB-PSO-LS

To account for our association method in the COMB-PSO framework, we consider the classification performance of the current global best solution (*gbest*) when redundant features are removed (LS_gbest). As highlighted in the flowchart of COMB-PSO-LS in Fig. 3 *gbest* is replaced by LS_gbest, if LS_gbest shows better performance than *gbest*, while *gbest* remains otherwise. As a consequence, our local search method allows the escape from local optima and enhances the stability of results. However, a challenge remains to find the right point to start local search as computational costs are significantly increased if local search is triggered too soon. In turn, strongly relevant features risk to be discounted in the selected subsets if local search is started too late. Fine tuning identified solutions, we limit local searches to the exploitation phase that is separated from exploitation phase through a sigmoid function ω_S returning the dynamic inertia weight ω (see Eq. 6 and Fig. 2). Specifically, we approximate the transition point between exploration and exploitation modes as the iteration number τ when both linear and sigmoid functions intersect (i.e. $\omega_S(\tau) = \omega_L(\tau), 0 < \tau \ll T$), and allow local search when iteration $t > \tau$. However, local search during the exploitation phase does not guarantee a fine tuned final solution. In fact, given the multi-objective optimization nature of the fitness function (for detail see subsection 3.3), local search solutions (LS_gbest) may be ignored compared to more cost effective gbest solutions during the iteration process of the PSO. Therefore,

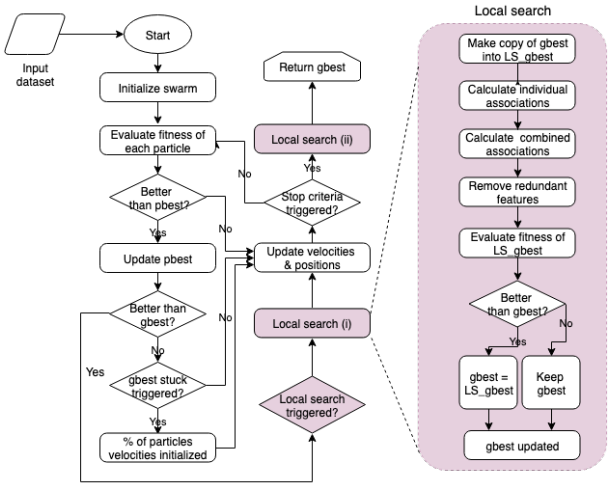

Figure 3: Flowchart of the COMB-PSO-LS algorithm, highlighting local search steps. The right box presents details of the local search, where redundant features of the globally best solution (*gbest*) are removed in the LS_gbest solution. LS_gbest is assigned gbest if its classification performance is enhanced.

we trigger local search after the stopping criteria of the iteration process has been met, as depicted by the box labeled "Local search (ii)" in the flowchart of our method (Fig. 3).

3.3 Problem Representation

Applying COMB-PSO-LS to different datasets, we formulate the objective functions as a weighted sum problem by

$$\min \quad \mathfrak{F} = \alpha E_F + (1-\alpha)\frac{|F|}{|D|}. \tag{10}$$

Specifically, F is the selected feature subset, D is the set of all features in the whole dataset, E_F is a the classification error rate, and α is a weight factor balancing the importance of the number of features and classification performance. As proposed in [36], $\alpha \in [0.6, 0.9]$, we set $\alpha = 0.8$.

Since COMB-PSO-LS is a wrapper approach, we utilize Random Forest (RF), an ensemble classification algorithm, to evaluate the classification accuracy of the selected features [11]. In particular, we use $mtry = \sqrt{n}$, $ntree = 5,000$, $nodesize = 1$, where n is the number of features, $mtry$ is the number of input variables tried at each split, $ntree$ is the number of trees in each forest and $nodesize$ is the minimum size of the terminal nodes.

During the search process, we randomly sample 70% as the training set and 30% as the test set and employ 10-fold cross-validation to evaluate the classification accuracy of the selected feature subset on the training set. Finally, the selected features are evaluated on the test set to obtain testing classification accuracy.

4 EXPERIMENTAL RESULTS

To test the performance of COMBPSO-LS we create synthetic datasets, establishing strongly relevant, weakly relevant and irrelevant features. Furthermore, we apply our algorithm to three cancer related gene expression datasets with a large number of genes and a limited amount of samples (Table 1) .

4.1 Experimental Datasets

We utilize the last of the three Monks datasets [33] that have 6 discrete features $\{f_0, \cdots, f_5\}$. Class labels are 1 if ($f_3 = 1$ and $f_4 = 3$) or ($f_4 \neq 4$ and $f_1 \neq 3$). Specifically, the Monks dataset has no redundant features, and the most important feature subset is $\{f_1, f_3, f_4\}$. The two other synthetic datasets have 4 continuous features. In the Synthetic 1 set, features f_2 and f_3 are copies of the first two features ($f_2 = f_0, f_3 = f_1$). The class label is set to 1 if the average of the first two features is greater than 0.5. Therefore, there are four optimal feature subsets in the Synthetic 1 dataset, $\{f_0, f_1\}, \{f_0, f_3\}, \{f_1, f_2\}$ or $\{f_2, f_3\}$. In the Synthetic 2 set, the first two features are random variables in $[0, 1]$. The 3^{rd} feature is the average of the first two, $f_2 = \frac{f_0+f_1}{2}$ while the 4^{th} feature is a copy of the first feature, $f_3 = f_0$. As a consequence, redundancy occurs in any feature subset that contains f_0 together with f_3 or (f_1 and f_2). The class label is determined by feature f_2. In particular, the class label is set to 1 if $f_2 > 0.5$, suggesting that the optimal feature subset of the Synthetic 2 dataset is $\{f_2\}$. In both the Synthetic 1 and Synthetic 2 sets, the remaining features are irrelevant with random values in [0,1]. To test the stability and the scalability of the different algorithms, we exponentially expand the three synthetic datasets up to 10^4 features by adding random values between [0,1] [41].

Table 1: Characteristics of synthetic and gene expression datasets

Dataset	# samples	# features	# classes
SYNTHETIC DATASETS			
Monks	432	10 000	2
Synthetic 1	200	10 000	2
Synthetic 2	200	10 000	2
GENE EXPRESSION DATASETS			
Leukemia	72	12 582	3
Prostate Tumor	95	16 535	2
B-cell lymphoma	77	6 428	2

As for real gene expression data (Table 1) we use a set of 72 *Leukemia* patient samples [2], consisting of 28 Acute Myeloid Leukemia (AML), 24 Acute Lymphoblastic Leukemia (ALL) and 20 Mixed-Lineage Leukemia (MLL) cases, capturing the expression levels of 12,582 genes. The *Prostate Tumor* [50] dataset provides expression levels of 16,535 genes in a total of 95 samples where 52 samples referred to tumor samples and the remainder to non-disease controls. The *Diffusive Large B-Cell Lymphoma* dataset [49] captures 58 patients with Diffuse Large B-Cell Lymphomas (DLBCL) and 14 patients with Follicular Lymphomas where each patient sample features 6,428 genes.

Table 2: Hyper parameters of the experimental set-up. Note that velocity boundaries in COMB-PSO and COMB-PSO-LS are not symmetric and are governed by Eq. 7.

Parameters	BPSO		COMB-PSO-LS	
	MIN	MAX	MIN	MAX
ω	0.4	0.9	0.4	0.9
c_1, c_2	2.05		1.7	2.1
(a, b) in Eq. 6			(.6, 8)	
velocity	-6.0	6.0	-6.0	0.25
λ in Eq. 7			1/32	
(θ, γ) in Subsec.2.5.4			(5, 20%)	
swarm size	300		300	
# iterations	3000		3000	

4.2 Experimental Setup

Finding the right parameter configuration for metaheuristic algorithms like PSO is a serious optimization problem, prompting the publication of many methods to automatically tune parameters to solve different problem instances [4, 21, 46, 56, 65]. Shi and Eberhart [48] analyzed the impact of inertia weight and maximum velocity on the performance of the PSO algorithm, providing guidelines for the selection of these parameters. Accounting for both simplicity and efficiency, linearly decreasing inertia weight is the most widely used setting, while $\omega_{max} = 0.9$ and $\omega_{min} = 0.4$ are accepted parameter choices. Furthermore, the relative values of acceleration coefficients c_1 and c_2 are critical, strongly impacting the performance of the underlying algorithm. When the value of the cognitive acceleration coefficient c_1 increases, the attraction of particles towards *pbest* is enhanced with the attraction towards *gbest* decreasing at the same time. In turn, increasing social acceleration coefficient c_2 compared to cognitive acceleration coefficient c_1 increases attraction of particles towards *gbest*. While values of c_1 and c_2 are generally kept constant, empirically best choices of c_1 and c_2 appear to be 2.05. In this work, and given the dynamic nature of c_1 and c_2 as introduced by COMB-PSO in subsection 2.5.2, we use a setting where c_1, c_2 vary between $c_{min} = 1.7$ and $c_{max} = 2.1$ and the transition coefficients are set to $a = .6$ and $b = 8$. Velocity boundaries are other factors that impact PSO performance. When boundaries are too large, particles move erratically and are swiftly attracted to *gbest* without thoroughly exploring the search space, increasing the risk of getting trapped in local optima. If boundaries are too narrow, movement of particles is excessively restricted, leading to computational overhead increases and inability of the algorithm to converge. For that reason, COMB-PSO introduced an asymmetric boundaries coefficient, as defined in Eq. 7, which is set empirically to $\lambda = 1/32$. To control premature convergence by avoiding stagnation, COMB-PSO introduced in subsection 2.5.4, both a stagnation coefficient θ which represents the number of iterations *gbest* being trapped before firing the turbulence operator and a turbulence coefficient $\gamma, \gamma \in [0, 1]$ which is the swarm fraction of the turbulence operator (i.e. the percentage of the swarm resetting their velocities). These two coefficients are set empirically to $\theta = 5$ and $\gamma = .2$. While no formal rule for the selection of the swarm size exists, rule of thumb stipulates that swarm sizes should be chosen proportional to the dimension of the underlying problem. Furthermore, swarm size impacts the performance of PSO as a smaller swarm leads to particles trapped in local optima while a larger swarm slows the performance of the algorithm. Swarm size is set to 300 particles and the number of iterations is set to 3000. All parameters are presented in Table 2.

4.3 Stability measures

To evaluate the stability of FSS algorithms, similarity measures are usually required to determine the divergence of selected subsets [28] as well as the convergence towards the optimal subsets. Usually, stability and classification accuracy are independent measures, suggesting that stability may not necessarily guarantee good classification results and *vice versa*. Considering both stability and classification accuracy for the performance evaluation of COMB-PSO-LS, we use two different stability measures. First, we apply a similarity metric when the optimal subset is unknown (*i.e.* for the gene expression datasets). Second, we consider a consistency metric when the optimal subset is known (*i.e.* for the synthetic datasets).

4.3.1 Similarity metric. Dune et al. [13] introduced a similarity-based metric as the average of all pairwise stability measures of solutions defined as

$$\overline{\mathcal{SM}}_M = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathcal{SM}(F_i, F_j). \tag{11}$$

$\overline{\mathcal{SM}}_M \in [0,1]$, where 0 indicates an empty intersection between all pairs of subsets F_i, F_j, while 1 points to the observation that all subsets of the system are identical. Furthermore, $\mathcal{SM}$ is the underlying similarity metric, while M is the number of trials. Kuncheva et al. [31], introduced a similarity metric that captures the correlation between features by

$$\mathcal{SM}(F_i, F_j) = \frac{|F_i \cap F_j| + \mathcal{SD}(F_i, F_j)}{|F_i \cup F_j| + 1}, \tag{12}$$

where $\mathcal{SD}$ is a statistical dependence measure between two subsets (For examples, see [3, 5, 16, 17, 22, 34, 44, 45, 52, 53]). As a consequence, $\mathcal{SM} \in [0,1]$, suggesting that $\mathcal{SM} = 0$ when two subsets have no intersection and no association. In turn, $\mathcal{SM} = 1$ when the two subsets are equal. However, two subsets with no intersection may still have a value greater than 0 if correlated features exist.

4.3.2 Consistency metric. The consistent identification of optimal subsets, defined as the smallest subsets that contain all strongly relevant features, is of particular interest to high-dimensional FSS problems. Since prior knowledge of the strongly relevant features and the optimal subsets exists considering synthetic datasets, we introduce a consistency metric which accounts for relevance, irrelevance and redundancy of selected features. In particular, we consider $\mathbf{F} = \{F_1, \cdots, F_r\}$ as the set of r selected subsets and F^* as the set of optimal solutions that provide only relevant and non-redundant features. To determine the overlap of $\mathbf{F}$ and F^* we define the consistency score $CM(\mathbf{F}) \in [0, 1]$ that reflects the degree of matching between the obtained set of solutions and known optimal solutions by

$$\mathcal{CM}(\mathbf{F}) = \frac{R_F}{R_F + R'_F + I_F} \tag{13}$$

where R_F, R'_F and I_F are the number of solutions containing relevant, redundant and irrelevant features, respectively. In particular, both R_F and R'_F are incremented if a solution has both relevant and redundant features. While $S(\mathbf{A}) = 0$ when no solution has a relevant feature, $\mathcal{CM}(\mathbf{F}) = 1$ when all solutions contain only relevant features (*i.e.* $F_i = F^*, \forall i \in [1, r]$). As a consequence, $\mathcal{CM}(\mathbf{F})$ penalizes (i) incomplete solutions, where relevant features are absent in $\mathbf{F}$, (ii) redundant solutions where more than enough relevant features appear in F_i and (iii) incorrect solutions, where irrelevant features occur in F_i.

4.4 Results

Here, we investigate the performance of COMB-PSO-LS in comparison to the standard BPSO and COMB-PSO, by collecting 30 solutions for each dataset from 30 independent runs. In particular, we assess obtained feature sets by measuring the mean sizes $\langle FS \rangle$ and mean classification error $\langle \%E \rangle$ of feature sets as well as introduce two other measures that capture the subsets propensity to provide strongly relevant features and their stability with the optimal subset as well as the similarity between the subsets.

Table 3 suggests that feature subsets obtained by applying our novel variant COMB-PSO-LS to synthetic datasets are (i) small and (ii) allow high classification accuracy. Furthermore, feature sets (iii) largely capture strongly relevant features and are (iv) highly similar, strongly outperforming BPSO and COMB-PSO.

While synthetic datasets provide a large number of features and samples, our gene expression datasets are high-dimensional as well but have a low number of samples. Applying our algorithms to the three different cancer gene expression datasets, we evaluate their performance by measuring the mean size of selected gene subsets $\langle FS \rangle$, the corresponding mean classification error rate $\langle \%E \rangle$ and determine the similarity of obtained gene subsets $\langle SM \rangle$. In the absence of ground truth (*i.e.* strongly relevant genes and optimal subsets), we therefore cannot establish the stability $\langle SM \rangle$ of obtained gene sets. In comparison to BPSO and COMB-PSO, COMB-PSO-LS provides the smallest gene subsets, that have a higher rate of similarity. Such observations indicate that the application of COMB-PSO-LS on gene expression datasets potentially allows us to find a small set of biomarkers in the underlying disease that distinguish between disease and control cases.

Although still low, gene subsets obtained with COMB-PSO-LS provide higher mean error rates $\langle \%ER \rangle$ compared to COMB-PSO. As global search feeds the local search operator, we surmise that swarm particles are limited in their propensity to thoroughly cover the search space therefore missing relevant features. However, the combination of all gene subsets that we obtain by applying COMB-PSO-LS to disease specific gene expression data allow us to observe a drop of the mean error rate to $< 2\%$.

Although the rates of similarity of gene subsets obtained with COMB-PSO-LS are higher compared to COMB-PSO and BPSO, they remain nevertheless moderate. Such an observation is potentially a consequence of the underlying data, as many relevant genes with highly correlated expression exist. Although relevant features are preserved during local search, our approach does not guarantee that all relevant genes can be covered in a single small subset, pointing to the presence of subsets with genes that are equivalent in their propensity to distinguish between disease and control cases.

To determine statistical significance of BPSO and COMB-PSO-LS results, we apply pairwise t-tests, allowing us to observe significant differences of distributions of subset sizes and classification error (Table 4). Although all p-values are significantly less than the confidence threshold $p < 0.05$ t-test values on the other hand, show a large difference in terms of subset size in favor of COMB-PSO and a relatively small difference in terms of classification error in favor of BPSO (negative t-test values). Nevertheless, we can conclude that the proposed method COMB-PSO-LS makes an impressive improvement over BPSO in terms of selected subset size while keeping classification error in a close range.

5 CONCLUSION

Given high-dimensional datasets with tens of thousands of features and few hundreds samples, we introduced COMB-PSO-LS that allowed us to find smallest and stable feature subsets, eliminating irrelevant and redundant features. In particular, we integrated a local search strategy to the framework of the previously introduced COMB-PSO algorithm based on the Randomized Dependency Coefficient that is known for its efficiency with non linear correlations. The combination of the association-based local search that guides the search process of the particle swarm and the global search strategy led to small, similar feature subsets that allowed high classification accuracy and captured the most salient features, significantly outperforming previous approaches. Applying our approach to gene expression datasets, we found small sets of genes that distinguished cancer from control cases while their similarity was limited. We expect that a multi-objective optimization representation of the objective function that embeds a measure of similarity in the evaluation of the Pareto front will allow us to increase the similarity and classification accuracy of gene subsets.

REFERENCES

[1] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. 2009. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26, 3 (2009), 392–398.

[2] Scott A. Armstrong, Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. den Boer, Mark D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub, and Stanley J. Korsmeyer. 2001. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* 30 (03 Dec 2001), 41 EP –. http://dx.doi.org/10.1038/ng765 Article.

[3] Francis R Bach and Michael I Jordan. 2002. Kernel independent component analysis. *Journal of machine learning research* 3, Jul (2002), 1–48.

[4] Indrajit Bhattacharya and Shukla Samanta. 2010. Parameter selection and performance study in particle swarm optimization. In *AIP Conference Proceedings*, Vol. 1298. AIP, 564–570.

[5] Leo Breiman and Jerome H Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80, 391 (1985), 580–598.

[6] Suresh Dara Chandra Sekhara Rao Annavarapu and Haider Banka. 2016. Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. *EXCLI journal* 15 (2016), 460.

[7] Li-Yeh Chuang, Hsueh-Wei Chang, Chung-Jui Tu, and Cheng-Hong Yang. 2008. Improved binary PSO for feature selection using gene expression data. *Computational Biology and Chemistry* 32, 1 (2008), 29–38.

[8] Li-Yeh Chuang, Sheng-Wei Tsai, and Cheng-Hong Yang. 2011. Improved binary particle swarm optimization using catfish effect for feature selection. *Expert Systems with Applications* 38, 10 (2011), 12699–12707.

[9] Li-Yeh Chuang, Cheng-Hong Yang, and Jung-Chike Li. 2011. Chaotic maps based on binary particle swarm optimization for feature selection. *Applied Soft*

Table 3: Performance results using both synthetic datasets (MONKS, Synthetic 1 and 2) and gene expression datasets (Leukemia, Prostate tumor and B-Cell Lymphoma). Comparing results obtained with BPSO, COMB-PSO and COMB-PSO-LS, we measure the mean sizes of feature sets $\langle FS \rangle$, mean classification error $\langle \%E \rangle$, as well as the consistency measure $\langle CM \rangle$ and the similarity measure $\langle SM \rangle$ of obtained feature sets F. Best performance results are indicated in Bold.

Dataset types →	Synthetic			Gene expression		
Algorithms →	BPSO	COMB-PSO	COMB-PSO-LS	BPSO	COMB-PSO	COMB-PSO-LS
Benchmarks ↓	**MONKS**			**LEUKEMIA**		
$\langle FS \rangle$	5,056	25	**2**	1,615	23	**2**
$\langle \%E \rangle$	5.6%	5.4%	**2.79%**	1.0%	**0.0%**	3.8%
$\langle CM \rangle$	6×10^{-4}	.007	**.98**	NA	NA	NA
$\langle SM \rangle$	.07	.002	**.91**	.06	.00083	**.12**
	SYNTHETIC 1			**PROSTATE TUMOR**		
$\langle FS \rangle$	5,209	249	**2**	2,114	7	**2**
$\langle \%E \rangle$	5.8%	13.3%	**0.0%**	**0.0%**	**0.0%**	2.7%
$\langle CM \rangle$	5×10^{-4}	.008	**1.0**	NA	NA	NA
$\langle SM \rangle$	.085	.0078	**.67**	.07	.00019	**.15**
	SYNTHETIC 2			**B-CELL LYMPHOMA**		
$\langle FS \rangle$	5,032	90	**1**	823	9	**3**
$\langle \%E \rangle$	5.6%	0.1%	**0.0%**	1.5%	**0.0%**	2.5%
$\langle CM \rangle$	7×10^{-4}	.01	**1.0**	NA	NA	NA
$\langle SM \rangle$	.07	.009	**1.0**	.08	.00081	**.23**

Table 4: Paired sample t-tests comparing sizes and classification error of feature subsets obtained with BPSO and COMB-PSO-LS

Dataset	Subset Size		Class. Error	
	t-test	p-value	t-test	p-value
LEUKEMIA	157.3	8.9×10^{-67}	-7.5	1.4×10^{-9}
PROSTATE TUMOR	258.1	1.8×10^{-90}	-7.1	3.8×10^{-6}
B-CELL LYMPHOMA	124.8	4.7×10^{-82}	-3.1	2.9×10^{-7}
MONKS	357.9	5.6×10^{-72}	7.9	2.7×10^{-8}
SYNTHETIC 1	372.5	6.2×10^{-66}	15.6	4.8×10^{-6}
SYNTHETIC 2	345.8	7.4×10^{-68}	14.9	3.9×10^{-7}

Computing 11, 1 (2011), 239–248.

[10] Hassen Dhrif, Luis G. Sanchez Giraldo, Miroslav Kubat, and Stefan Wuchty. 2019. A Stable Combinatorial Particle Swarm Optimization for Scalable Feature Selection in Gene Expression Data. *arXiv e-prints*, Article arXiv:1901.08619 (Jan. 2019), arXiv:1901.08619 pages. arXiv:cs.NE/1901.08619

[11] Ramón Díaz-Uriarte and Sara Alvarez De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* 7, 1 (2006), 3.

[12] Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology* 3, 02 (2005), 185–205.

[13] Kevin Dunne, Padraig Cunningham, and Francisco Azuaje. 2002. Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research* (2002), 1–22.

[14] RC Eberhart and J Kennedy. 1995. Particle swarm optimization, proceeding of IEEE International Conference on Neural Network. *Perth, Australia* (1995), 1942–1948.

[15] Artur J Ferreira and MáRio AT Figueiredo. 2012. An unsupervised approach to feature discretization and selection. *Pattern Recognition* 45, 9 (2012), 3048–3060.

[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, Mar (2012), 723–773.

[17] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[18] Fei Han, Wei Sun, and Qing-Hua Ling. 2014. A novel strategy for gene selection of microarray data based on gene-to-class sensitivity information. *PloS one* 9, 5 (2014), e97530.

[19] Fei Han, Chun Yang, Ya-Qi Wu, Jian-Sheng Zhu, Qing-Hua Ling, Yu-Qing Song, and De-Shuang Huang. 2017. A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-Class Sensitivity Information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 14, 1 (2017), 85–96.

[20] Min Han and Weijie Ren. 2015. Global mutual information-based feature selection approach using single-objective and multi-objective optimization. *Neurocomputing* 168 (2015), 47–54.

[21] Ali B Hashemi and Mohammad Reza Meybodi. 2011. A note on the learning automata based algorithms for adaptive parameter selection in PSO. *Applied Soft Computing* 11, 1 (2011), 689–705.

[22] Trevor J Hastie. 2017. Generalized additive models. In *Statistical models in S*. Routledge, 249–307.

[23] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one* 6, 12 (2011), e28210.

[24] Zengyou He and Weichuan Yu. 2010. Stable feature selection for biomarker discovery. *Computational biology and chemistry* 34, 4 (2010), 215–225.

[25] Zena M Hira and Duncan F Gillies. 2015. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics* 2015 (2015).

[26] H Hannah Inbarani, Ahmad Taher Azar, and G Jothi. 2014. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer methods and programs in biomedicine* 113, 1 (2014), 175–185.

[27] George H John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*. Elsevier,

121–129.
[28] Alexandros Kalousis, Julien Prados, and Melanie Hilario. 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems* 12, 1 (2007), 95–116.
[29] J. Kennedy and R. Eberhart. 1995. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, Vol. 4. 1942–1948 vol.4. https://doi.org/10.1109/ICNN.1995.488968
[30] J. Kennedy and R. C. Eberhart. 1997. A discrete binary version of the particle swarm algorithm. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Vol. 5. 4104–4108 vol.5. https://doi.org/10.1109/ICSMC.1997.637339
[31] Ludmila I Kuncheva. 2007. A stability index for feature selection.. In *Artificial intelligence and applications*. 421–427.
[32] Carmen Lai, Marcel JT Reinders, and Lodewyk Wessels. 2006. Random subspace method for multivariate feature selection. *Pattern recognition letters* 27, 10 (2006), 1067–1076.
[33] Moshe Lichman et al. 2013. UCI machine learning repository. (2013).
[34] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. 2013. The randomized dependence coefficient. In *Advances in neural information processing systems*. 1–9.
[35] Jonathan L Lustgarten, Vanathi Gopalakrishnan, and Shyam Visweswaran. 2009. Measuring stability of feature selection in biomedical datasets. In *AMIA annual symposium proceedings*, Vol. 2009. American Medical Informatics Association, 406.
[36] Mohd Saberi Mohamad, Sigeru Omatu, Safaai Deris, and Michifumi Yoshioka. 2011. A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. *IEEE Transactions on Information Technology in Biomedicine* 15, 6 (2011), 813–822.
[37] Jacqueline Moore and Richard Chapman. 1999. Application of particle swarm to multiobjective optimization. *Department of Computer Science and Software Engineering, Auburn University* 32 (1999).
[38] Parham Moradi and Mozhgan Gholampour. 2016. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing* 43 (2016), 117–130.
[39] Parham Moradi and Mehrdad Rostami. 2015. A graph theoretic approach for unsupervised feature selection. *Engineering Applications of Artificial Intelligence* 44 (2015), 33–45.
[40] Parham Moradi and Mehrdad Rostami. 2015. Integration of graph clustering with ant colony optimization for feature selection. *Knowledge-Based Systems* 84 (2015), 144–161.
[41] Bach Hoai Nguyen. 2018. Evolutionary Computation for Feature Selection in Classification. (2018).
[42] Sarah Nogueira and Gavin Brown. 2015. Measuring the stability of feature selection with applications to ensemble methods. In *International Workshop on Multiple Classifier Systems*. Springer, 135–146.
[43] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.
[44] Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. 2012. Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682* (2012).
[45] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. 2011. Detecting novel associations in large data sets. *science* 334, 6062 (2011), 1518–1524.
[46] A Rezaee Jordehi and Jasronita Jasni. 2013. Parameter selection in particle swarm optimisation: a survey. *Journal of Experimental & Theoretical Artificial Intelligence* 25, 4 (2013), 527–542.
[47] Q. Shen, W.M. Shi, and W. Kong. 2008. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Comput. Biol. Chem.* 32 (2008), 53–60.
[48] Yuhui Shi and Russell C Eberhart. 1998. Parameter selection in particle swarm optimization. In *International conference on evolutionary programming*. Springer, 591–600.
[49] Margaret A. Shipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C. T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8 (01 Jan 2002), 68 EP –. http://dx.doi.org/10.1038/nm0102-68 Article.
[50] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 2 (2002), 203 – 209. https://doi.org/10.1016/S1535-6108(02)00030-2
[51] Petr Somol and Jana Novovicova. 2010. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 11 (2010), 1921–1939.
[52] Gábor J Székely and Maria L Rizzo. 2009. Brownian distance covariance. *The annals of applied statistics* (2009), 1236–1265.
[53] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35, 6 (2007), 2769–2794.
[54] Sina Tabakhi and Parham Moradi. 2015. Relevance–redundancy feature selection based on ant colony optimization. *Pattern recognition* 48, 9 (2015), 2798–2811.
[55] Sina Tabakhi, Parham Moradi, and Fardin Akhlaghian. 2014. An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence* 32 (2014), 112–123.
[56] Ioan Cristian Trelea. 2003. The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information processing letters* 85, 6 (2003), 317–325.
[57] Alper Unler, Alper Murat, and Ratna Babu Chinnam. 2011. mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences* 181, 20 (2011), 4625–4641.
[58] Susana M Vieira, Luís F Mendonça, Goncalo J Farinha, and João MC Sousa. 2013. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Applied Soft Computing* 13, 8 (2013), 3494–3504.
[59] Zhichun Wang, Minqiang Li, and Juanzi Li. 2015. A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure. *Information Sciences* 307 (2015), 73–88.
[60] Momiao Xiong, Xiangzhong Fang, and Jinying Zhao. 2001. Biomarker identification by feature wrappers. *Genome Research* 11, 11 (2001), 1878–1887.
[61] Bing Xue, Liam Cervante, Lin Shang, Will N Browne, and Mengjie Zhang. 2012. A multi-objective particle swarm optimisation for filter-based feature selection in classification problems. *Connection Science* 24, 2-3 (2012), 91–116.
[62] Bing Xue, Mengjie Zhang, and Will N Browne. 2013. Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE transactions on cybernetics* 43, 6 (2013), 1656–1671.
[63] Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang. 2008. A Hybrid Feature Selection Method for Microarray Classification. *IAENG International Journal of Computer Science* 35, 3 (2008).
[64] Feng Yang and KZ Mao. 2011. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8, 4 (2011), 1080–1092.
[65] Fei Ye. 2017. Particle swarm optimization-based automatic parameter selection for deep neural networks and its applications in large-scale and high-dimensional data. *PloS one* 12, 12 (2017), e0188746.
[66] Zhang Yong, Gong Dun-wei, and Zhang Wan-qiu. 2016. Feature selection of unreliable data using an improved multi-objective PSO algorithm. *Neurocomputing* 171 (2016), 1281–1290.
[67] Lei Yu and Huan Liu. 2004. Redundancy based feature selection for microarray data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 737–742.
[68] Yong Zhang, Dun-wei Gong, and Jian Cheng. 2017. Multi-objective particle swarm optimization approach for cost-based feature selection in classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 14, 1 (2017), 64–75.
[69] Yudong Zhang, Shuihua Wang, Preetha Phillips, and Genlin Ji. 2014. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems* 64 (2014), 22–31.