A class-specific metaheuristic technique for explainable relevant feature selection.

EZENKWU, C.P., AKPAN, U.I. and STEPHEN, B.U.-A.

2021

Machine Learning with Applications 6 (2021) 100142

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

A class-specific metaheuristic technique for explainable relevant feature selection

Chinedu Pascal Ezenkwu [a,c,*], Uduak Idio Akpan [b,c], Bliss Utibe-Abasi Stephen [a]

[a] Electrical/Electronic & Computer Engineering Department, University of Uyo, Nigeria
[b] Electrical/Electronic Engineering Department, Akwa Ibom State University, Nigeria
[c] School of Engineering, University of Aberdeen, United Kingdom

ARTICLE INFO

Keywords:
Feature selection
Explainable AI
XAI
Genetic algorithms
Metaheuristics
Nature-inspired

ABSTRACT

A significant amount of previous research into feature selection has been aimed at developing methods that can derive variables that are relevant to an entire dataset. Although these approaches have revealed substantial improvements in classification accuracy, they have failed to address the problem of explainability of outputs. This paper seeks to address this problem of identifying explainable features using a class-specific feature selection method based on genetic algorithms and the one-vs-all strategy. Our proposed method finds relevant features for each class in the dataset and uses these features to enable more accurate classification, and also interpretation of the outputs. The results of our experiments demonstrate that the proposed method provides descriptive insights into prediction outputs, and also outperforms popular global feature selection techniques in the classifications of high dimensional and noisy datasets. Since there are no known challenging benchmark datasets for evaluating class-specific feature selection algorithms, this paper also recommends an approach for combining disparate datasets for this purpose.

1. Introduction

The aims of feature selection are to identify a subset of high-dimensional features that can improve the predictive accuracy of a classifier, minimise computation time (Pourpanah et al., 2019) and/or enable the interpretability of the result (Pourpanah et al., 2019). The explainability of prediction outputs is vital for the real-world application of machine learning in domains such as medicine, aerospace and finance where rationales for a model's decisions are *desiderata* for a user's trust (Došilović et al., 2018). Even though global feature selection methods have demonstrated that they can improve the predictive power of classification algorithms, they are limited in the number of insights and level of understanding that they can provide a human inspector with. These methods seek to derive relevant features for an entire dataset. This is problematic because these features cannot be linked to any specific prediction output, and therefore they cannot provide transparency to the process. A possible solution to this issue is to employ a class-specific or local feature selection technique. In this method, relevant features are identified for each class in a dataset, making it possible for a human inspector to interpret the rationale for each predicted class by linking the decision back to those features identified for the class. This process of associating a set of attributes to a prediction outcome is an explainable AI (XAI) technique called feature attribution (Janzing et al., 2020; Liu & Avci, 2019).

In this paper, we present a class-specific feature selection method based on a metaheuristic optimisation technique (Glover & Kochenberger, 2006; Talbi, 2009) and the one-vs-all (Rifkin & Klautau, 2004) strategy. Due to the exponential amount of time required to find the best subset of features in a high-dimensional dataset, a feature selection problem can be described as an NP-hard problem (Žerovnik, 2015). Metaheuristic algorithms such as genetic algorithms (GA) (Whitley, 1994), simulated annealing (Van Laarhoven & Aarts, 1987), tabu search (Glover, 1989), particle swarm optimisation (Kennedy & Eberhart, 1995) and so on, are generally identified as the most plausible techniques for combinatorial optimisation problems (Yagiura & Ibaraki, 2001). Although a number of different metaheuristic algorithms can be suitable for a feature selection problem, this paper adopts GA because it is intuitive and naturally copes very well with discrete optimisation tasks. The one-vs-all strategy transforms a multiclass problem into multiple binary classification problems, making it possible to select the relevant features for a specific class using GA. With a few of these selected features, it is easier to realise the decision rules for each class using an intuitive or whitebox algorithm such as the decision tree.

In the existing literature, a feature selection technique is often evaluated on the basis of its accuracy with respect to the identified relevant features (Chandrashekar & Sahin, 2014; Tang & He, 2016). We argue that this evaluation metric is not sufficient for class-specific

* Corresponding author at: Electrical/Electronic & Computer Engineering Department, University of Uyo, Nigeria.
E-mail addresses: chineduezenkwu@uniuyo.edu.ng (C.P. Ezenkwu), uduakidio@aksu.edu.ng (U.I. Akpan), blissustephen@gmail.com (B.U.-A. Stephen).

[1] https://community.fico.com/s/explainable-machine-learning-challenge.

https://doi.org/10.1016/j.mlwa.2021.100142
Received 9 December 2020; Received in revised form 13 August 2021; Accepted 13 August 2021
Available online 26 August 2021

feature selection tasks. In addition to prediction performance, a class-specific feature selection method requires some domain knowledge for the interpretation of the feature attributions. To the best of our knowledge, there are no challenging benchmark datasets for a class-specific feature selection problem. For example, the FICO community hosts an anonymised Home Equity Line of Credit (HELOC) dataset for its explainable machine learning challenge.[1] This dataset has twenty three well-defined credit behaviour features from a credit bureau. However, the dataset is not appropriate for a class-specific feature selection challenge because it is a binary class dataset. In a two-class problem, every useful feature is expected to contribute unique information to the classification problem (Zhang et al., 2019) and must have a minimum branching factor of two for the two classes. We agree that global relevant features are adequate for explaining either class in a binary class dataset. For example, if age is an important variable for offering university admission, then age is likely to be an important factor for denying selection. The only difference between the two classes are the decision rules that apply to them based on this variable. As a result of the aforementioned observation, this paper also presents an approach for combining disparate datasets for more challenging class-specific feature selection problems.

The remainder of this paper proceeds as follows: Section 2 presents background information; Section 3 reviews related papers while the proposed method is explained in Section 4. Experiments and results are presented in Section 5 and Section 6 respectively, and Section 7 concludes the paper.

2. Background information

2.1. Feature selection methods

These are generally categorised as filter, wrapper and embedded methods (Gnana et al., 2016; Pereira et al., 2018). This section reviews these methods.

2.1.1. Filter methods

Filter methods generally use variable ranking techniques as criteria for feature ordering and selection for classification or regression tasks. Feature relevancy in filter methods is generally based on the correlation between predictor variables and targets. Features which are independent of the class values are irrelevant (Law et al., 2004) and should be discarded.

One of the most popular criteria for scoring feature relevance in filter-based feature selection methods is the Pearson correlation coefficient (Battiti, 1994; Guyon & Elisseeff, 2003) presented in Eq. (1).

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \tag{1}$$

where x_i is the ith feature, Y is the class label, $cov()$ is the covariance and $var()$ is the variance. This criterion can only detect the linear dependence between the variable and the target.

An information theoretic ranking criterion such as mutual information (MI) also serves as a measure of dependency between two variables (Battiti, 1994; Guyon & Elisseeff, 2003; Lazar et al., 2012; Zhang et al., 2019). The MI between X and Y is given by:

$$I(Y, X) = H(Y) - H(Y|X) \tag{2}$$

where $H(Y)$ is Shannon's entropy, defined as follows:

$$H(Y) = -\sum_{y} p(y) log(p(y)) \tag{3}$$

and $H(Y|X)$ is the conditional entropy of output Y given that a variable X is observed. $H(Y|X)$ is defined in Eq. (4).

$$H(Y|X) = -\sum_{x}\sum_{y} p(x, y) log(p(y|x)) \tag{4}$$

Eq. (2) shows that if Y and X are independent, then MI will be zero otherwise MI is greater than zero.

Other common filter-based feature selection methods including relief (Kira & Rendell, 1992) and reliefF (Kononenko, 1994) which are based on nearest neighbours. ReliefF is the multiclass variant of the relief algorithm.

2.1.2. Wrapper methods

Wrapper methods use a search technique in identifying a subset of features that will optimise the performance measure of a certain classifier. This performance measure or objective function is dependent on the type of problem. For example, a regression evaluation criterion can be R-squared while classification evaluation criteria can be accuracy, recall, precision, f1-score and so on. Common search algorithms used for wrapper feature selection include the branch and bound method (Kohavi et al., 1997; Narendra & Fukunaga, 1977) and several metaheuristic algorithms. In addition to the heuristic search algorithms, some wrapper feature selection methods are based on sequential selection algorithms such as the sequential forward selection (SFS), sequential backward selection (SBS), sequential forward floating Selection (SFFS) and sequential backward floating selection (SBFS) (Chandrashekar & Sahin, 2014; Dunne et al., 2002; Ferri et al., 1994; Somol et al., 1999). These methods iteratively add or remove features until a termination criterion is met.

2.1.3. Embedded methods

While filter methods are independent of any induction algorithm, wrapper methods use a classifier to evaluate the quality of feature subsets. Moreover, wrapper methods do not "incorporate knowledge about the specific structure a classification or regression function and can therefore be combined with any learning machine" (Lal et al., 2006). Embedded methods differ from these two methods because they seek to incorporate a feature selection capacity in a learning algorithm.

For example, Guyon et al. present an embedded method that uses the weights of a classifier for feature ranking (Guyon & Elisseeff, 2003; Guyon et al., 2002). Weight w_j is defined as follows:

$$w_j = \frac{\mu_j(+) - \mu_j(-)}{\sigma_j(+) + \sigma_j(-)} \tag{5}$$

where μ_j and σ_j are the mean and standard deviation of the samples in class (+) and class (-). Large positive w_j values indicate strong correlation with class (+) whereas large negative w_j values indicate strong correlation with class (-).

The feature ranking presented in Eq. (5) can be used to design a classifier as follows:

$$D(\mathbf{x}) = \mathbf{w}.(\mathbf{x} - \mu) \tag{6}$$

where $\mathbf{w}$ is the rank of the features or weight, defined in Eq. (5) and μ is the mean of the data - $\mu = (\mu(+) + \mu(-))/2$.

Several embedded methods involve a change in the objective function of a classifier in order to learn the feature ranking using the model weight vector (Guyon & Elisseeff, 2003; Guyon et al., 2002). For example, the support vector machine (SVM) (Suthaharan, 2016) cost function was modified to perform recursive feature elimination (RFE). This method is known as the SVM-RFE method (Boser et al., 1992; Guyon et al., 2002; Mundra & Rajapakse, 2009). A similar technique has been developed for a multilayer neural network (Setiono & Liu, 1997).

2.2. Genetic algorithms

GA is a metaheuristic search technique based on Charles Darwin's principle of natural selection (Genlin, 2004; Thengade & Dondal, 2012). As is typical with all population-based algorithms, GA starts with a randomly generated population of candidate solutions which iteratively improve from one generation to the next. The selection of individuals

into the next generation depends on a fitness or objective function. In addition to selection, GA employs other biologically inspired operators such as mutation and crossover for generating high-quality individuals. There are several ways of applying these operators depending on whether we have a real-valued (Corcoran & Sen, 1994; Wu et al., 2007) or a binary-valued (Cao et al., 2005; Pampara et al., 2006) optimisation problem. In this paper, feature selection is considered to be a binary optimisation problem. Algorithm 1 summarises GA procedure.

Algorithm 1: Genetic Algorithm pseudocode

```
START;
Generate initial population;
Compute fitness;
while Termination Criterion := FALSE do
    Selection;
    Crossover;
    Mutation;
    Compute fitness;
end
```

2.3. One-vs-all strategy

In multiclass classification, one-vs-all (ova) strategy involves training a single classifier for each class, with the samples of that class labelled as 1 and all other samples as 0 (Bishop, 2006). Given a dataset $\{X_i, y_i\}_{i=1}^N$ and a classifier F, where $y_i \in \{1, \dots, K\}$, one-vs-all strategy aims to train a list of binary classifiers $F_k\ \forall k \in \{1, \dots, K\}$ such that for each k the dataset is transformed to $\{X_i, z_i^k\}_{i=1}^N$ where $z_i^k = 1$ if $y_i = k$ and $z_i^k = 0$ otherwise. To predict a new input vector $\mathbf{x}$, k of the classifier F_k that gives the highest confidence score is reported.

3. Related work

Using GA as the feature selection technique, Maleki et al. improved the performance of a k-Nearest Neighbours (kNN) algorithm in detecting the early stages of lung cancer from 99.80% to 100% accuracy (Maleki et al., 2020). Li et al. present a feature selection method that combines feature weighted kNN and the real-valued GA algorithm in ranking features in a high dimensional dataset (Li et al., 2020). A new variant of GA named as the binary chaotic genetic algorithm (BCGA) showed an improvement over traditional GA in feature selection tasks using AMIGOS (A Dataset for Affect, Personality and Mood Research on Individuals and Groups) and two healthcare datasets having large feature space (Tahir et al., 2020).

Paniri et al. have proposed a multi-label feature selection method using swarm intelligence ant colony optimisation (ACO) (Paniri et al., 2020). The method has shown a better performance over five state-of-the-art feature selection algorithms in nine well-known datasets using the multi-level kNN (ML-kNN) classifier. ACO has also been applied as a feature selection method in financial crisis prediction (Uthayakumar et al., 2020), in breast cancer detection (Saranya & Sasikala, 2020) and in the assessment of humorous speeches by TED speakers (Adi et al., 2020).

Further examples of metaheuristic techniques have also been applied to feature selection problems. For example, the hyper learning binary dragonfly algorithm (HLBDA), a dragonfly-based method, has demonstrated an improved performance in the classification of twenty-one datasets from the University of California Irvine (UCI) repository (Dua & Graff, 2017) and Arizona State University, together with a coronavirus disease (COVID-19) dataset as presented by Too and Mirjalili (2020). Simulated annealing has been used as a feature selection method in flash-flood hazard assessment (Hosseini et al., 2020) while the particle swarm optimisation(PSO) technique (Khan, 2020) and a hybrid method using PSO and the flower pollination algorithm (Tawhid & Ibrahim, 2020) have been used as feature selection techniques for software effort prediction and on popular UCI datasets respectively.

Gao et al. propose the dynamic change of selected feature with the class (DCSF) method (Gao et al., 2018), a feature selection technique based on class-specific mutual information variation. Unlike the traditional feature selection methods, DCSF considers the dynamic change of selected features with the class. However, instead of a set of relevant features for each class, DCSF yields global features for the entire dataset.

Although these methods have performed well in different areas of their applications, none of them have considered a class-specific feature selection task, making their outcomes more difficult to interpret by a human inspector. While to the best of our knowledge, no metaheuristic algorithm has been applied to class-specific feature selection, there are a limited number of non-metaheuristic algorithms aimed at class-specific feature selection tasks (Ruan et al., 2020; Yuan et al., 2020). A number of these works measured the performance of the proposed methods in terms of their prediction accuracies. As a result, some popular datasets published on the UCI repository (Dua & Graff, 2017) have been used for evaluating these methods. We argue that since these datasets are not designed for class-specific feature selection tasks, they are not suitable for evaluating class-specific feature selection methods.

The work in this paper is entailed by these limitations. This paper proposes a metaheuristic class-specific feature selection technique based on GA. In order to address the difficulties imposed by the lack of benchmark datasets for evaluating class-specific feature selection methods, this paper also presents an approach for combining well-known datasets for more challenging class-specific feature selection tasks.

4. Methodology

We define class-specific feature selection in the context of a multiclass supervised classification task. Given a dataset $\{X_i, y_i\}_{i=1}^N$ and a binary classifier F, where $y_i \in \{1, \dots, K\}$ and $X_i \in R^d$; and $d \geq 2$, a class-specific feature selection algorithm $\mathcal{G}$ seeks to identify any set of features $S_k \subseteq \{f_j\}_{j=1}^d$ (where $\{f_j\}_{j=1}^d$ is a set of all d features) for each class k such that the performance score of the kth class classifier F_k is maximised for each k using $\{X_i^k, z_i^k\}_{i=1}^N$ where each feature in X^k belongs to S_k and $z_i^k = 1$ if $y_i = k$ and 0 otherwise.

Mathematically,

$$S_k^* = \underset{S_k \subseteq \{f_j\}_{j=1}^d}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \mathfrak{I}(F_k(X_i^k) := z_i^k) \tag{7}$$

$\forall k \in \{1, \dots, K\}$, where $\mathfrak{I}(.)$ is an indicator function that returns a 1 when F_k is correct and 0 otherwise and $:=$ is a comparison operator.

While the one-vs-all strategy converts any multiclass classification problem to multiple binary classification tasks, GA searches for the relevant features for each binary classification task so that the selected features are capable of describing the class of interest for each of the binary classification tasks.

Next, we summarise the steps for the proposed GA-based class-specific feature selection algorithm as follows:

Step I: One-Vs-All process – For each class $k \in \{1, \dots, K\}$, $\{X_i, z_i^k\}_{i=1}^N \leftarrow \textbf{\textit{OneVsAll}}(\{X_i, y_i\}_{i=1}^N)$ where ***OneVsAll***() is a function that transforms the multiclass dataset $\{X_i, y_i\}_{i=1}^N$ into a binary class dataset $\{X_i, z_i^k\}_{i=1}^N$ such that $z_i^k = 1$ if $y_i = k$ and 0 otherwise.

The result of this process is a set of all the K binary class datasets, $\mathfrak{D} = \{(X_i, z_i^k)_{i=1}^N\}_{k=1}^K$.

Note that the one-vs-all strategy often lead to a class imbalance problem when used with long multiclass datasets. Due to this, we apply random oversampling of the minority class before the feature selection step.

Step II: Feature selection for each class – For each $\{X_i, z_i^k\}_{i=1}^N \in \mathfrak{D}$ do the following:

Table 1
Experimental datasets.

Dataset	Dimension	Number of classes	Citation of relevant paper(s)
Ionosphere	351 by 34	2	Sigillito et al. (1989)
Glass identification	214 by 10	7	Evett and Ernest (1987)
Dermatology	366 by 33	6	Güvenir et al. (1998)
Isolet	7797 by 617	26	Fanty and Cole (1991), Dietterich and Bakiri (1991)
Ionosphere+glass	565 by 44	9	Ionosphere $\oplus$ glass identification datasets using the method described in Fig. 2
Statlog heart	270 by 13	2	Brown (2004)
Landsat satellite	6435 by 36	6	Cheng et al. (2018)
Semeion handwritten digit	1593 by 256	10	Buscema (1998)
Soybean	35 by 47	4	Michalski (1980)
Splice-junction gene	3175 by 60	3	Noordewier et al. (1991)

Table 2
Hyperparameters for the SVM.

Regulariser C	Kernel	Decision function shape	Learning rate γ
1.0	Linear	One-vs-rest	$\frac{1}{Number of features}$

Table 3
Hyperparameters for the Random Forest RF.

Number of trees in the forest	Maximum depth	Minimum samples split	Split measurement criterion
100	2	2	Gini

Table 4
Hyperparameters for the GA.

Fitness function	Population size	Crossover probability	Mutation probability	Tournament size
Accuracy	200	0.5	0.2	3

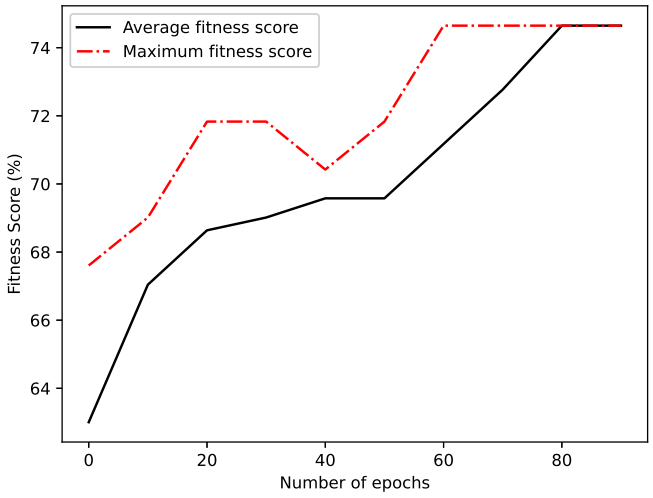


Fig. 1. GA-ova learning curve on the glass identification data using SVM with the maximum of five selected features.

A: Generate initial population – A population $\mathcal{P}$ of M candidate solutions, in which each solution is a list of d binary values i.e. $\{0,1\}^d$, is generated. The values 0 and 1 indicate if a feature in the corresponding position is selected or not. 1 means that the feature is selected while 0 means otherwise.

B: Compute fitness – For each candidate solution $c \in \mathcal{P}$, a binary classifier F_k^c is trained on $\{X_i, z_i^k\}_{i=1}^N$ and the performance score(fitness) $\mathcal{J}_k^c$ is calculated.

C: Selection – A set Q of fittest individuals are selected based on their fitness scores, $\mathcal{J}_k^c$.

D: Crossover – The candidate solutions in Q are selected in pairs and recombined to generate new population of candidate solutions (Holland, 1975).

E: Mutation – With a small probability an arbitrary bit in a candidate solution is flipped so as to avoid local minima as well as to maintain genetic diversity from one generation to another.

F: Termination – If the change in the average performance scores between generations is above a threshold (1.0×10^{-6} in this case) return to B, otherwise terminate.

For clarity and brevity, the method proposed in this paper has been referred to throughout as GA-ova, an acronym for one-vs-all genetic algorithms.

5. Experiments

5.1. Descriptions

The aims of our experiments are to evaluate the predictive power of a classifier when combined with GA-ova, and to also assess the ability of the method to select relevant features for each class in a dataset. To demonstrate that GA-ova is able to improve the predictive power of a classifier, its performances on different datasets are compared with MI, RFE, DCSF and a GA-based method without one-vs-all technique. The performance of each feature selection algorithm is evaluated using SVM and Random Forest (RF) as classifiers. Fig. 1 shows the learning curve of the GA-ova on the glass identification dataset using SVM with the maximum of five selected features. The average fitness score of the population converges to the maximum fitness score at the 80th epoch.

5.2. Experimental datasets

Due to computing and time constraints, we evaluate the performance accuracy of GA-ova using only ten different datasets. To evaluate the ability of GA-ova to select relevant features for each class, we combined two disparate datasets using the approach illustrated in Fig. 2. In Fig. 2 two datasets, (X_A, y_A) and (X_B, y_B), are combined with a Gaussian noise of 0 mean and standard deviation of 1 i.e. $\epsilon \sim \mathcal{N}(0,1)$. It is expected that an effective class-specific feature selection method will successfully attribute variables in the d part of the features if any class in y_A is predicted and for any class in y_B the selected features will belong in the k part of the features.

Table 1 summarises the datasets used in the experiments. All the datasets except the ionosphere+glass dataset are published on the UCI machine learning repository (Dua & Graff, 2017). The ionosphere+glass dataset is synthesised following the approach presented in Fig. 2. The

Table 5
Performance accuracies of different feature selection algorithms on the datasets in Table 1. In each row, the value of the best performing technique is presented in bold typeface.

Dataset	Max. no of features	Accuracy % (Mean ± Standard error)											
		GA		MI		RFE		F-test		DCSF		GA-ova	
		SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Ionosphere	1	80.17 ± 1.32	81.20 ± 0.56	75.57 ± 0.94	76.63 ± 1.32	77.00 ± 1.02	83.4 ± 0.59	76.3 ± 2.03	76.93 ± 1.86	66.40 ± 1.89	81.90 ± 1.15	**84.07** ± 0.68	81.13 ± 0.64
	3	89.07 ± 1.18	89.63 ± 0.41	79.63 ± 0.83	82.87 ± 0.98	88.30 ± 1.55	89.80 ± 1.00	82.00 ± 0.26	89.00 ± 1.769	81.90 ± 1.97	84.40 ± 1.17	**90.63** ± 0.91	90.07 ± 0.19
	5	90.10 ± 0.23	90.30 ± 1.03	80.13 ± 1.42	85.67 ± 0.97	85.53 ± 1.39	**91.37** ± 0.60	82.43 ± 2.71	89.13 ± 1.72	80.10 ± 1.92	86.20 ± 1.14	88.67 ± 0.41	90.67 ± 0.90
Glass Identification	1	52.57 ± 0.90	58.07 ± 0.97	48.27 ± 0.97	54.00 ± 0.87	52.73 ± 0.79	58.00 ± 1.02	44.10 ± 0.95	45.67 ± 0.83	39.33 ± 0.88	40.00 ± 2.00	56.87 ± 0.83	**59.70** ± 0.87
	3	65.13 ± 0.10	64.1 ± 0.75	51.07 ± 0.97	62.60 ± 0.81	60.97 ± 0.97	**66.73** ± 0.85	55.20 ± 1.02	63.73 ± 0.92	45.00 ± 1.15	53.33 ± 1.53	58.03 ± 0.61	59.83 ± 0.79
	5	64.97 ± 1.13	63.97 ± 0.83	54.00 ± 0.87	65.1 ± 1.02	68.7 ± 0.64	63.83 ± 0.92	59.63 ± 0.92	62.5 ± 0.87	49.00 ± 1.15	55.00 ± 2.00	61.1 ± 0.87	**70.73** ± 0.79
Dermatology	1	50.67 ± 0.53	50.13 ± 0.53	51.20 ± 1.15	50.87 ± 0.88	25.60 ± 1.15	49.60 ± 1.15	49.60 ± 1.13	49.60 ± 0.00	45.40 ± 1.14	45.4 ± 0.93	81.00 ± 1.14	**85.10** ± 1.91
	3	81.27 ± 0.27	78.80 ± 2.20	63.6 ± 1.15	76.23 ± 0.90	72.70 ± 1.15	66.90 ± 1.15	49.60 ± 1.14	50.13 ± 0.27	61.30 ± 1.13	59.70 ± 0.00	**94.77** ± 0.23	93.13 ± 0.53
	5	91.43 ± 0.27	88.47 ± 0.83	76.37 ± 1.18	76.00 ± 1.15	90.90 ± 1.15	76.00 ± 1.15	53.70 ± 1.16	54.50 ± 0.00	78.20 ± 1.12	63.90 ± 1.15	**95.00** ± 0.49	94.20 ± 1.15
Isolet	1	9.00 ± 0.87	22.83 ± 0.99	9.03 ± 0.99	9.20 ± 1.02	10.30 ± 0.94	7.80 ± 0.87	11.10 ± 0.87	11.73 ± 0.85	10.10 ± 0.00	7.80 ± 0.00	29.70 ± 0.81	**37.70** ± 0.64
	3	28.10 ± 0.84	22.83 ± 0.85	19.10 ± 0.87	16.70 ± 0.53	29.87 ± 0.97	23.17 ± 1.05	16.00 ± 1.02	18.97 ± 0.97	26.60 ± 0.00	18.30 ± 0.00	61.93 ± 0.99	**62.20** ± 1.02
	5	44.77 ± 0.89	38.03 ± 0.99	20.93 ± 0.85	16.87 ± 0.97	44.93 ± 0.92	30.17 ± 0.89	26.97 ± 0.97	21.10 ± 0.87	28.30 ± 0.00	23.5 ± 0.00	**68.63** ± 0.79	66.50 ± 0.87
Ionosphere + Glass	1	52.13 ± 0.34	51.83 ± 0.85	50.17 ± 0.89	46.30 ± 0.95	48.40 ± 1.77	51.5 ± 0.52	50.03 ± 0.99	49.97 ± 1.59	64.03 ± 1.13	56.10 ± 1.16	51.87 ± 0.82	**62.70** ± 0.75
	3	60.23 ± 0.99	62.00 ± 0.87	51.20 ± 0.95	52.23 ± 0.92	59.83 ± 1.59	53.87 ± 1.35	51.50 ± 1.29	59.23 ± 0.92	64.03 ± 1.13	55.00 ± 1.13	68.87 ± 0.89	**72.10** ± 1.02
	5	67.13 ± 0.99	69.17 ± 1.04	58.10 ± 1.02	63.10 ± 0.87	67.17 ± 1.68	63.00 ± 0.87	58.23 ± 0.92	62.70 ± 0.75	66.84 ± 1.15	56.60 ± 1.14	71.03 ± 0.92	**73.27** ± 2.04
Statlog heart	1	57.83 ± 0.62	77.81 ± 0.23	**78.91** ± 1.10	66.13 ± 0.73	67.82 ± 1.03	77.70 ± 1.03	71.10 ± 1.13	71.10 ± 1.12	75.61 ± 0.94	76.70 ± 0.93	57.80 ± 1.14	59.82 ± 1.01
	3	72.97 ± 0.07	77.80 ± 1.01	79.53 ± 1.11	71.12 ± 0.70	75.64 ± 0.95	78.90 ± 1.15	79.53 ± 0.86	77.72 ± 0.58	76.70 ± 1.15	78.91 ± 0.67	**80.00** ± 0.27	76.45 ± 0.98
	5	79.27 ± 0.34	77.83 ± 0.36	79.27 ± 0.98	83.85 ± 1.01	81.13 ± 1.00	77.82 ± 0.45	79.27 ± 0.78	**85.61** ± 0.23	75.60 ± 0.88	74.43 ± 1.02	78.91 ± 0.59	81.82 ± 0.45
Landsat satellite	1	55.12 ± 0.65	51.52 ± 0.59	55.83 ± 0.14	50.34 ± 1.08	51.35 ± 0.82	51.56 ± 0.61	55.80 ± 0.43	50.00 ± 0.39	55.82 ± 0.96	50.34 ± 0.83	**71.13** ± 0.23	69.62 ± 0.84
	3	**77.80** ± 0.23	71.23 ± 0.34	76.15 ± 0.67	76.23 ± 0.52	77.02 ± 0.24	52.37 ± 0.85	76.12 ± 0.15	64.82 ± 0.97	76.32 ± 0.65	52.47 ± 0.86	77.02 ± 0.69	76.30 ± 1.02
	5	80.1 ± 0.52	70.4 ± 0.41	77.34 ± 0.55	77.32 ± 0.47	79.91 ± 0.23	67.90 ± 0.91	77.42 ± 0.67	65.52 ± 0.52	78.14 ± 0.12	52.43 ± 0.20	**83.13** ± 0.79	79.71 ± 0.23
Semeion handwritten digit	1	17.91 ± 0.12	17.32 ± 0.23	17.54 ± 0.39	17.40 ± 0.72	18.72 ± 0.34	16.71 ± 0.36	15.63 ± 0.23	15.12 ± 0.25	17.54 ± 0.20	17.42 ± 0.05	**27.80** ± 0.51	20.93 ± 0.34
	3	25.72 ± 0.23	28.50 ± 0.15	25.32 ± 0.38	24.42 ± 0.49	**40.32** ± 0.17	31.72 ± 0.55	29.81 ± 0.72	29.01 ± 0.37	36.81 ± 0.12	35.00 ± 0.24	39.51 ± 0.23	31.71 ± 0.24
	5	43.71 ± 0.34	39.72 ± 0.41	36.51 ± 0.45	34.24 ± 0.17	**46.41** ± 0.21	31.62 ± 0.47	30.04 ± 0.43	33.82 ± 0.68	40.31 ± 0.23	35.42 ± 0.33	45.12 ± 0.34	39.71 ± 0.59
Soybean	1	68.51 ± 0.46	68.75 ± 0.29	87.51 ± 0.23	87.50 ± 0.56	68.92 ± 0.36	68.81 ± 0.37	50.10 ± 0.45	51.23 ± 0.34	87.53 ± 0.19	87.52 ± 0.26	**100.00** ± 0.00	62.52 ± 0.54
	3	**100.00** ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00	93.81 ± 0.23	93.81 ± 0.23	87.53 ± 0.19	87.53 ± 0.19	**100.00** ± 0.00	**100.00** ± 0.00
	5	**100.00** ± 0.00	**100.00** ± 0.00	93.83 ± 0.13	**100.00** ± 0.00	**100.00** ± 0.00	68.83 ± 0.62	93.81 ± 0.23	93.81 ± 0.23	87.53 ± 0.19	87.53 ± 0.19	**100.00** ± 0.00	93.80 ± 0.84
Splice-junction gene	1	54.52 ± 0.34	57.13 ± 0.23	62.83 ± 0.45	61.43 ± 0.44	63.12 ± 0.34	62.83 ± 0.24	62.83 ± 0.23	62.83 ± 0.23	62.80 ± 0.85	62.80 ± 0.85	57.92 ± 0.32	**64.12** ± 0.35
	3	67.92 ± 0.24	74.23 ± 0.36	76.11 ± 0.31	73.21 ± 0.67	76.35 ± 0.40	77.13 ± 0.45	77.18 ± 0.34	77.10 ± 0.32	72.12 ± 0.37	65.13 ± 0.49	77.10 ± 0.32	**78.53** ± 0.12
	5	72.30 ± 0.23	78.59 ± 0.23	83.78 ± 0.95	86.82 ± 0.49	87.23 ± 0.34	85.61 ± 0.38	87.23 ± 0.36	86.73 ± 1.02	72.83 ± 0.19	71.42 ± 0.29	**87.41** ± 0.71	78.59 ± 0.67

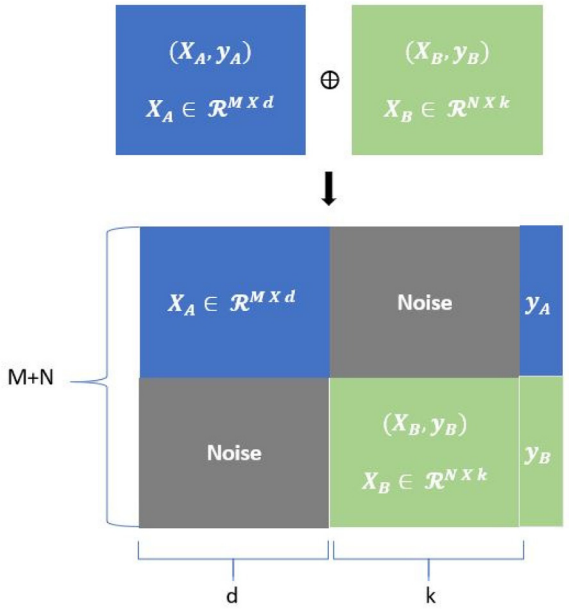


Fig. 2. Combining disparate dataset for class-specific feature selection evaluation.

only data cleaning process we applied to each of the datasets was to remove rows that contain empty or null values.

5.3. Hyperparameters

In this section, we present the hyperparameters used during the experiments. Tables 2–4 are the hyperparameters for the SVM, RF and GA respectively. To justify the comparisons of the competing feature selection methods, these hyperparameters are the same for all the methods in which the corresponding algorithm is applicable.

6. Results and discussions

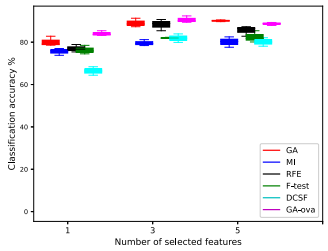
6.1. Classification performance

The performances of the different feature selection techniques were evaluated using the datasets presented in Table 1 with SVM and the Random Forest classifiers. Figs. 3–4 and Table 5 demonstrate the results of the experiment. In each row in Table 5, the value of the best performing technique is presented in bold typeface. Fig. 5 presents the average classification performance of the feature selection algorithms across all datasets.

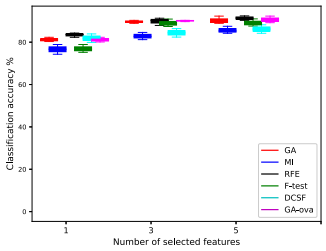
From the results of the experiment, it can be inferred that GA-ova does not produce a better performance than the global feature selection techniques in certain situations. For example, Figs. 3(a) and 3(b) presents the comparative performances of the different methods in the classification of the ionosphere dataset. As described in Table 1, the ionosphere dataset is a binary-class dataset, and as such, the one-vs-all strategy in GA-ova makes little or no difference to the performance of the method. The same situation applies to the Statlog heart dataset in Figs. 4(a) and 4(b)

Similarly, from Figs. 3(c) and 3(d), the performances of GA-ova on the glass-identification dataset do not stand out amongst those of other methods. Although this dataset is a multiclass dataset of 7 classes, it consists of only 10 features. We can conclude that problems involving fewer features do not pose an adequate challenge to the global feature selection methods. While it is important to provide the definition of *fewer features* in this context, this is beyond the scope of this paper.

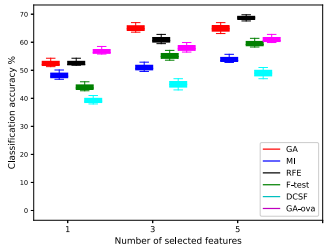
However, the GA-ova method has been shown to perform substantially better than other methods in Figs. 3(e)–3(j), 4(c)–4(d) and 4(h)–4(i). Figs. 3(g) and 3(h) show the performances of all the methods on the isolet dataset, which from Table 1 consists of 26 classes and 617 features. While the classification of this dataset poses a significant



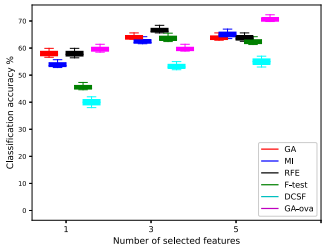
(a) Ionosphere dataset using SVM



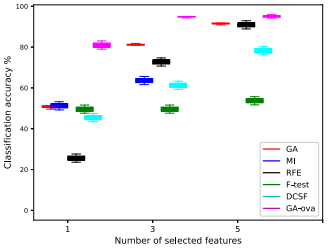
(b) Ionosphere dataset using Random Forest



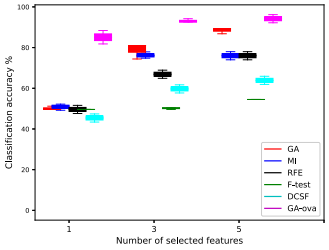
(c) Glass identification dataset using SVM



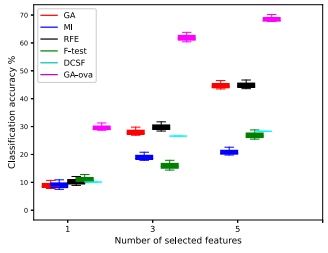
(d) Glass identification dataset using Random Forest



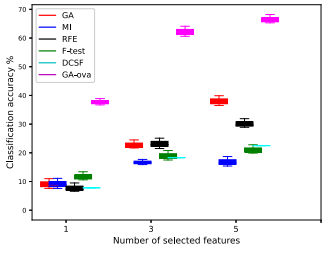
(e) Dermatology dataset using SVM



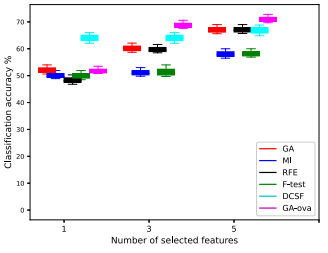
(f) Dermatology dataset using Random Forest



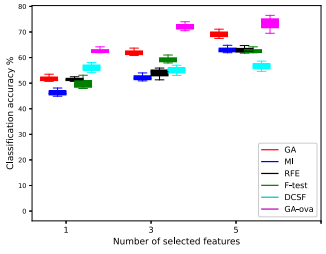
(g) Isolet dataset using SVM



(h) Isolet dataset using Random Forest



(i) Ionosphere+Glass dataset using SVM



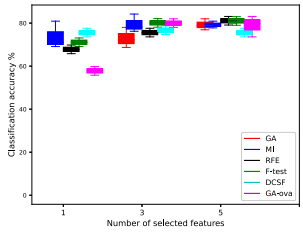
(j) Ionosphere+Glass dataset using Random Forest

Fig. 3. Performance accuracies of different feature selection algorithms on the datasets in Table 1.

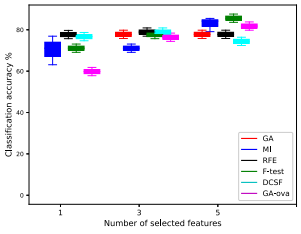
difficulty to the other methods, GA-ova demonstrated a remarkable performance on the dataset especially when the maximum number of selected features is three or more. From Figs. 3(i) and 3(j) GA-ova performed better than the other methods. As already described in Section 5.2, the ionosphere+glass dataset consists of disparate data with some Gaussian noise. GA-ova has been demonstrated to have higher classification accuracy than the other methods on this dataset. This is because unlike the global feature selection algorithms, GA-ova, a local feature selection method, identifies features which are relevant for each of the classes in the dataset. With the one-vs-all strategy, GA-ova did relatively well in discarding features that are likely to add some noise in the prediction of any of the classes.

6.1.1. *Friedman statistical hypothesis tests*

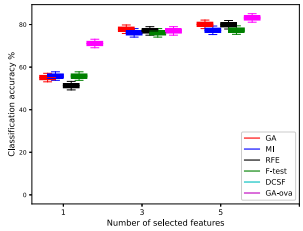
The Friedman test is a nonparametric equivalent of repeated measures analysis of variance (ANOVA) for comparing more than two



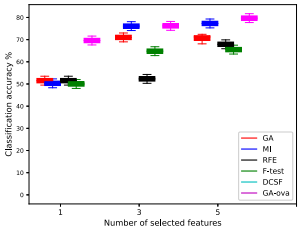
(a) Statlog heart dataset using SVM



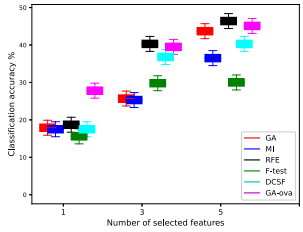
(b) Statlog heart dataset using Random Forest



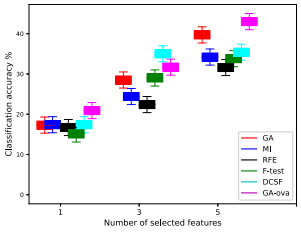
(c) landsat satellite dataset using Random Forest using SVM



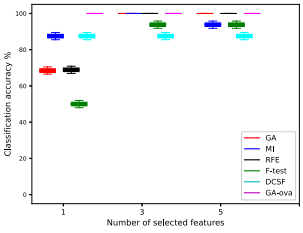
(d) landsat satellite dataset using Random Forest



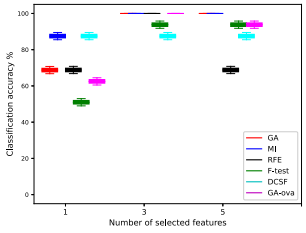
(e) Semeion handwritten digit dataset using SVM



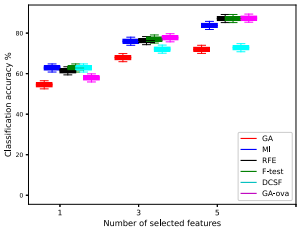
(f) Semeion handwritten digit dataset using Random Forest



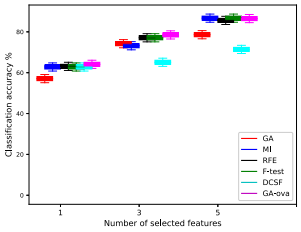
(g) Soybean dataset using SVM



(h) Soybean dataset using Random Forest



(i) Splice-junction gene dataset using SVM



(j) Splice-junction gene dataset using Random Forest

Fig. 4. Performance accuracies of different feature selection algorithms on the datasets in Table 1.

samples that are related. It gives a significant result, if at least one of the samples is different from the other samples (Zimmerman & Zumbo, 1993).

Fail to Reject H0: Paired distribution of performance scores across datasets are equal.

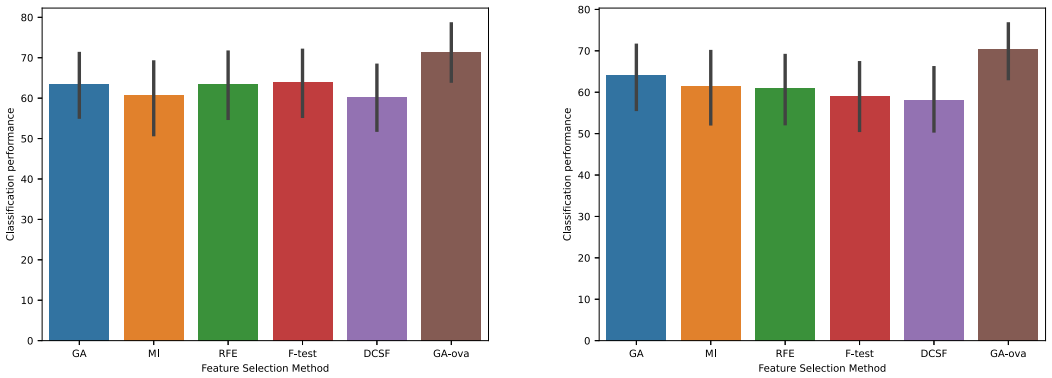
Reject H0: Paired distribution of performance scores across datasets are not equal.

For the performance scores of the feature selection algorithms due to SVM and Random Forest, Friedman test gives p-values of 5.5856e−08 and 1.6894e−05 respectively.

These small values of p indicate that at least one set of performance scores has a different distribution.

6.2. Identification of class-specific features

In addition to improving the predictive power of a classifier, another key potential of GA-ova is its capacity to select class-specific features for



(a) Evaluation with SVM (b) Evaluation with Random Forest

Fig. 5. Average performance of the feature selection algorithms across all datasets.

the interpretability of prediction outcomes. This potential is currently lacking in global feature selection methods as shown in Table 6. Table 6 presents the selected set of five features due to each feature selection method when applied to the ionosphere+glass dataset using SVM. As illustrated in Section 5.2, the ionosphere+glass dataset consists of disparate datasets with some Gaussian noise. The first part of the dataset consists of features f1 to f34 relevant for classes 1 and 2, while features f35 to f44 present some Gaussian noise to these classes. The second part of the dataset contains features f35 to f44 with classes 3–5,7–9 and features f1 to f34 are noise to these classes.

From Table 6, it can be observed that for each class in the ionosphere+glass dataset GA-ova performed fairly well in selecting features from the parts of features that belong to the original dataset. With the exception of class 8, GA-ova was able to select at least 60% of the features for each class from the parts of features that are associated with the class. However, we observed that some of the odd features that GA-ova associated to some of the classes, even though they are noise, have some descriptive characteristics and are useful for the classification of the associated classes. For example, Fig. 6 demonstrates that the odd feature f39 that is attributed to class 1 (i.e. data points in red) can contribute effectively to the classification of that class. From the Figure, f39 is able to identify substantial members of other classes (i.e. data points in blues). Combining f39 and the other features can provide a more powerful descriptor for classifying class 1. For example, Fig. 7 shows that the combination of f39 and f3 is able to separate out some class 1 data points from the cluster in the lower left corner of Fig. 6. In Fig. 8, a decision tree is used to demonstrate the interactions of the features selected for class 1 (f3,f7,f8,f18 and f39) in distinguishing this class from others. It can be seen that only f39 is sufficient for classifying 251 members of other classes correctly.

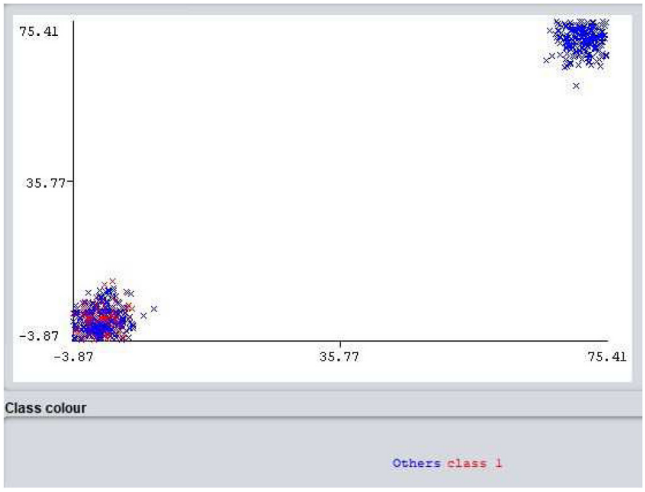


Fig. 6. The distribution of class 1 and other classes with respect to f39 in the ionosphere+glass dataset. Class 1 is in red while others is represented in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

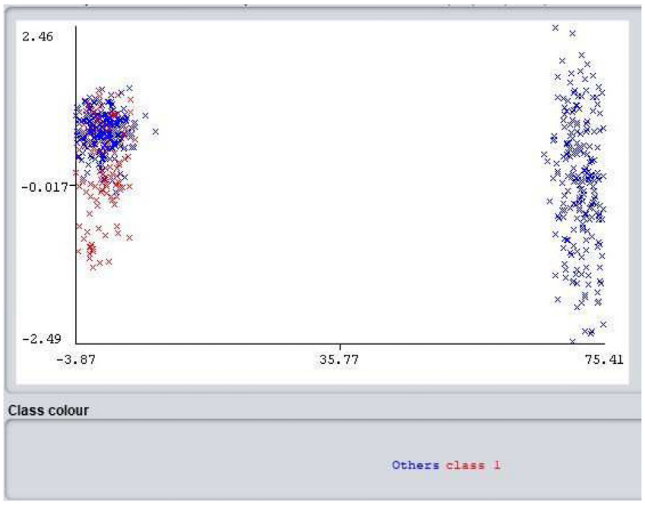


Fig. 7. The distribution of class 1 and other classes due to the interaction of f3 and f39 in the ionosphere+glass dataset. Class 1 is in red while others are represented in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7. Conclusions

This paper presents a class-specific feature selection method based on GA and the one-vs-all strategy. A major limitation of global feature selection algorithms is that they search for a set of features that will optimise the predictive power of a learning algorithm. While these methods have shown a substantial improvement in the predictive powers of classifiers, they are not capable of addressing the problem of explainability of prediction outcomes. This paper is intended to develop a method that identifies relevant features for each class in a dataset. It shows that our proposed method outperforms popular global feature selection techniques in classifying high dimensional and noisy datasets. This paper also recommended a strategy for combining disparate datasets for evaluating class-specific feature selection tasks.

Table 6
A set of five features selected by each method from the ionosphere+glass dataset using SVM.

Feature selection technique	Class	Selected features
GA	All	f7,f8,f22,f37,f41
MI	All	f1,f36,f37,f39,f41
RFE	All	f35,f36,f37,f38,f40
F-test	All	f36,f37,f38,f39,f41
DCSF	All	f43,f7,f25,f13,f31
GA-ova	Class 1	f3,f7,f8,f18,f39
	Class 2	f1,f5,f8,f24,f39
	Class 3	f36,f37,f38,f41
	Class 4	f13,f16,f37,f40, f41
	Class 5	f4,8,f17,f37, f38
	Class 7	f10,f37, f39, f40
	Class 8	f2,f12, f14, f36, f42
	Class 9	f28,f36, f37, f40, f41

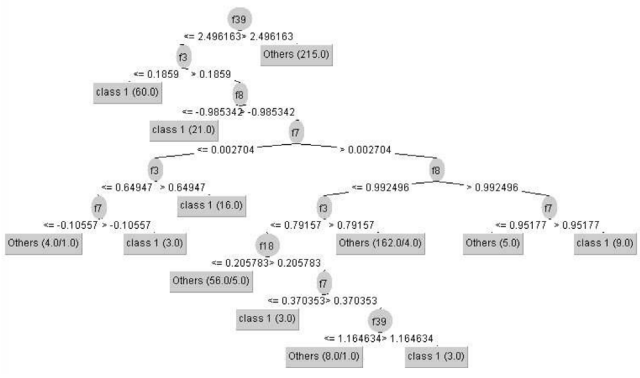


Fig. 8. A decision tree showing the interactions of the selected features for the prediction of class 1.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adi, D. P., Gumelar, A. B., Meisa, R. P. A., & Susilowati, S. (2020). Assessment of humorous speech by automatic heuristic-based feature selection. In *2020 International seminar on application for technology of information and communication (ISemantic)* (pp. 597–602). IEEE.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Brown, G. (2004). *Diversity in neural network ensembles* (Ph.D. thesis), Citeseer.

Buscema, M. (1998). Metanet*: The theory of independent judges. *Substance Use & Misuse*, *33*(2), 439–461.

Cao, J.-Y., Liang, J., & Cao, B.-G. (2005). Optimization of fractional order PID controllers based on genetic algorithms. In *2005 International conference on machine learning and cybernetics, Vol. 9* (pp. 5686–5689). IEEE.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28.

Cheng, L., Leung, A. C. S., & Ozawa, S. (2018). *vol. 11303, Neural information processing: 25th international conference, ICONIP 2018, Siem reap, Cambodia, December 13–16, 2018, Proceedings, Part III*. Springer.

Corcoran, A. L., & Sen, S. (1994). Using real-valued genetic algorithms to evolve rule sets for classification. In *Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence* (pp. 120–124). IEEE.

Dietterich, T. G., & Bakiri, G. (1991). Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *AAAI* (pp. 572–577). Citeseer.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.

Dunne, K., Cunningham, P., & Azuaje, F. (2002). Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research*, 1–22.

Evett, I. W., & Ernest, J. S. (1987). Rule induction in forensic science. Central research establishment. home office forensic science service. Aldermaston. *Reading*, Berkshire RG7 4PN.

Fanty, M., & Cole, R. (1991). Spoken letter recognition. In *Advances in neural information processing systems* (pp. 220–226).

Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. *vol. 16*, In *Machine intelligence and pattern recognition* (pp. 403–413). Elsevier.

Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. *Pattern Recognition*, *79*, 328–339.

Genlin, J. (2004). Survey on genetic algorithm [j]. *Computer Applications and Software*, *2*(1), 69–73.

Glover, F. (1989). Tabu search—part i. *ORSA Journal on Computing*, *1*(3), 190–206.

Glover, F. W., & Kochenberger, G. A. (2006). *vol. 57, Handbook of metaheuristics*. Springer Science & Business Media.

Gnana, D. A. A., Balamurugan, S. A. A., & Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, *975*, 8887.

Güvenir, H. A., Demiröz, G., & Ilter, N. (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, *13*(3), 147–165.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.

Holland, J. (1975). Adaptation in natural and artificial systems: an introductory analysis with application to biology. *Control and Artificial Intelligence*.

Hosseini, F. S., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H., & Haghighi, A. T. (2020). Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: application of the simulated annealing feature selection method. *Science of the Total Environment*, *711*, Article 135161.

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *International conference on artificial intelligence and statistics* (pp. 2907–2916). PMLR.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *vol. 4*, In *Proceedings of ICNN'95-international conference on neural networks* (pp. 1942–1948). IEEE.

Khan, M. Z. (2020). Particle swarm optimisation based feature selection for software effort prediction using supervised machine learning and ensemble methods: A comparative study. *Invertis Journal of Science & Technology*, *13*(1), 33–50.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249–256). Elsevier.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *European conference on machine learning* (pp. 171–182). Springer.

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Feature extraction* (pp. 137–165). Springer.

Law, M. H., Figueiredo, M. A., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1154–1166.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1106–1119.

Li, S., Zhang, K., Chen, Q., Wang, S., & Zhang, S. (2020). Feature selection for high dimensional data using weighted K-nearest neighbors and genetic algorithm. *IEEE Access*, *8*, 139512–139528.

Liu, F., & Avci, B. (2019). Incorporating priors with feature attribution on text classification. arXiv preprint arXiv:1906.08286.

Maleki, N., Zeinali, Y., & Niaki, S. T. A. (2020). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, *164*, Article 113981.

Michalski, R. S. (1980). Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of development an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, *4*(2), 125–161.

Mundra, P. A., & Rajapakse, J. C. (2009). SVM-RFE With MRMR filter for gene selection. *IEEE Transactions on Nanobioscience*, *9*(1), 31–37.

Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, (9), 917–922.

Noordewier, M. O., Towell, G. G., & Shavlik, J. W. (1991). Training knowledge-based neural networks to recognize genes in DNA sequences. In *Advances in neural information processing systems* (pp. 530–536).

Pampara, G., Engelbrecht, A. P., & Franken, N. (2006). Binary differential evolution. In *2006 IEEE international conference on evolutionary computation* (pp. 1873–1879). IEEE.

Paniri, M., Dowlatshahi, M. B., & Nezamabadi-pour, H. (2020). Mlaco: A multi-label feature selection algorithm based on ant colony optimization. *Knowledge-Based Systems*, *192*, Article 105285.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2018). Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, *49*(1), 57–78.

Pourpanah, F., Shi, Y., Lim, C. P., Hao, Q., & Tan, C. J. (2019). Feature selection based on brain storm optimization for data classification. *Applied Soft Computing*, *80*, 761–775.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, *5*(Jan), 101–141.

Ruan, S., Li, H., Li, C., & Song, K. (2020). Class-specific deep feature weighting for Naïve Bayes text classifiers. *IEEE Access*, *8*, 20151–20159.

Saranya, S., & Sasikala, S. (2020). Malignant breast cancer detection using feature selection and ant colony optimization deep learning technique. *Solid State Technology*, *63*(6), 3565–3580.

Setiono, R., & Liu, H. (1997). Neural-network feature selector. *IEEE Transactions on Neural Networks*, *8*(3), 654–662.

Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, *10*(3), 262–266.

Somol, P., Pudil, P., Novovičová, J., & Paclık, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, *20*(11–13), 1157–1163.

Suthaharan, S. (2016). Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification* (pp. 207–235). Springer.

Tahir, M., Tubaishat, A., Al-Obeidat, F., Shah, B., Halim, Z., & Waqas, M. (2020). A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare. *Neural Computing and Applications*, 1–22.

Talbi, E.-G. (2009). *vol. 74*, *Metaheuristics: From Design To Implementation*. John Wiley & Sons.

Tang, B., & He, H. (2016). FSMJ: Feature selection with maximum Jensen-Shannon divergence for text categorization. In *2016 12th world congress on intelligent control and automation (WCICA)* (pp. 3143–3148). IEEE.

Tawhid, M. A., & Ibrahim, A. M. (2020). Hybrid binary particle swarm optimization and flower pollination algorithm based on rough set approach for feature selection problem. In *Nature-inspired computation in data mining and machine learning* (pp. 249–273). Springer.

Thengade, A., & Dondal, R. (2012). Genetic algorithm-survey paper. In *MPGI national multi conference* (pp. 7–8). Citeseer.

Too, J., & Mirjalili, S. (2020). A hyper learning binary dragonfly algorithm for feature selection: A COVID-19 case study. *Knowledge-Based Systems*, Article 106553.

Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. (2020). Financial crisis prediction model using ant colony optimization. *International Journal of Information Management*, *50*, 538–556.

Van Laarhoven, P. J., & Aarts, E. H. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, *4*(2), 65–85.

Wu, C.-H., Tzeng, G.-H., Goo, Y.-J., & Fang, W.-C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications*, *32*(2), 397–408.

Yagiura, M., & Ibaraki, T. (2001). On metaheuristic algorithms for combinatorial optimization problems. *Systems and Computers in Japan*, *32*(3), 33–55.

Yuan, L.-m., Sun, Y., & Huang, G. (2020). Using class-specific feature selection for cancer detection with gene expression profile data of platelets. *Sensors*, *20*(5), 1528.

Žerovnik, J. (2015). Heuristics for NP-hard optimization problems-simpler is better!? *Logistics & Sustainable Transport*, *6*(1), 1–10.

Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. *Information Fusion*, *50*, 158–167.

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the wilcoxon test, the friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, *62*(1), 75–86.