Dimension Reduction in Classification using Particle Swarm Optimisation and Statistical Variable Grouping Information

Bing Xue[1], Mitchell C. Lane[1], Ivy Liu[2], and Mengjie Zhang[1]
[1]School of Engineering and Computer Science
[2]School of Mathematics and Statistics
Victoria University of Wellington, Wellington, New Zealand
Email: {Bing.Xue, Mitchell.Lane, Mengjie.Zhang}@ecs.vuw.ac.nz, ivy.liu@vuw.ac.nz

***Abstract*—Dimension reduction is a preprocessing step in many classification tasks, but reducing dimensionality and finding the optimal set of features or attributes are challenging because of the big search space and interactions between attributes. This paper proposes a new dimension reduction method by using a statistical variable grouping method that groups similar attributes into a group by considering interaction between attributes and using particle swarm optimisation as a search technique to adopt the discovered statistical grouping information to search optimal attribute subsets. Two types of approaches are developed, where the first aims to select one attribute from each group to reduce the dimensionality, and the second allows the selection of multiple attributes from one group to further improve the classification performance. Experiments on ten datasets of varying difficulties show that all the two approaches can successfully address dimension reduction tasks to decrease the number of attributes, and achieve the similar of better classification performance. The first approach selects a smaller number of attributes than the second approach while the second approach achieves better classification performance. The proposed new algorithms outperform other recent dimension reduction algorithms in terms of the classification performance, or further reduce the number of attributes while maintaining the classification performance.**

I. Introduction

In recent years, with the advances of data collection techniques, many real-world data mining tasks, e.g. classification, often include a large number of attributes (or variables or features). This causes the problem of "the curse of dimensionality" and leads to many issues, e.g. learning/classification algorithms fail to achieve satisfied accuracy, the classification process is time-consuming, and the trained classifier is too complicated to understand/interpret. Dimension reduction can address these issues by using only a small number of attributes. It can be achieved by attribute selection to select a smaller number of feature or feature construction to create a smaller set of new attributes [1], [2], [3], [4], [5]. Dimension reduction has been used in both supervised learning (e.g. classification that learns from *labelled data* to assign one of the predefined class labels to an instance), and unsupervised learning (e.g. clustering that learns from *unlabelled data* to group similar instances to different clusters), but the majority of the current work on dimension reduction is for classification. The focus of this work is dimension reduction via selecting a subset of relevant attributes in classification.

Existing dimension reduction algorithms can be classified into two categories: filter and wrapper approaches [1], [3]. Their main difference is whether a classification method is involved in the dimension reduction process. Wrappers uses a classification algorithm to measure the classification performance of the selected attributes to evaluate the selected attributes. Filters do not use any classification algorithm, which are often computationally cheaper and more general, but wrappers often can achieve better classification accuracy than filters [6].

Dimension reduction is challenging because of the *large search space* and the *interactions between variables*. The search space size is 2^n for a dataset with n variables. Many algorithms, such as greedy search based algorithms [6], have limitations, such as stagnation in local optimal or high computationally expensive. An efficient global search technique should be used to address dimension reduction problems. Furthermore, dimension reduction Evolutionary computation (EC) techniques include powerful "global" search algorithms and have been successfully applied to a variety of fields [7]. Among them, Particle swarm optimisation (PSO) is based on social intelligence, which has fewer parameters and is computationally cheaper than many other EC techniques, e.g. genetic algorithms (GAs) and genetic programming (GP) [7].

Feature interaction is a commonly appeared issue in classification tasks [6], [8]. Because of interactions between attributes, an individually relevant attribute may become less useful or redundant when combined with other attributes. On the other hand, a weakly relevant attribute could become very useful when used together with other attributes. In an "optimal" subset, attributes are expected to provide complementary information and can work together to increase the classification accuracy. Therefore, during dimension reduction, the removal or addition of attributes should consider the appearance or absence of other attributes, which increases the difficulty of dimension reduction tasks. Finding a way to cope with interactions between attributes is expected to increase the performance of a dimension reduction algorithm.

Meanwhile, interactions between attributes is also an important issue being considered in statistical data analysis. Applying statistical variable grouping methods [9], [10] through a model allows taking interactions between attributes into account to group relatively homogeneous attributes into groups (Note that *statistical variable grouping* methods aim to *group similar attributes* into a single group, which are different from the *clustering (unsupervised learning)* methods in data mining (e.g. K-means) that often aim to *cluster similar instances/examples* in the data set into one cluster [11]). Such statistical ideas could be useful to address interactions between attributes in dimension reduction, but very little work has been conducted in this area. Our recent work has shown that such statistical variable grouping information could be utilised to develop a good dimension reduction algorithm [12], [13], but they are just very preliminary work on this direction and further investigation is needed (the difference between this new work and the existing work [12], [13] are described in Section III).

Goals: The goal of this work is to propose a novel wrapper based dimension reduction approach in classification based on PSO and statistical variable grouping. To achieve this, a statistical variable grouping method is employed as a pre-processing step to group attributes into different groups. By utilising the statistical variable grouping information, two approaches are proposed to maximise the classification performance with the constraint of selecting one attribute from each group and allowing the selection of multiple attributes from each group, respectively. Specifically, we will investigate:

- whether the first new approach can select one attribute from each group to reduce the dimensionality, and obtain similar or better accuracy than using all the available attributes,
- whether the second new approach that allows selecting multiple attributes from a single group can further improve the classification performance,
- whether the newly developed approaches can outperform two conventional dimension reduction algorithms, a single objective PSO algorithm without using grouping information, and a method using grouping information.

II. Background

A. Particle Swarm Optimisation (PSO)

Particle swarm optimisation (PSO) [14] was proposed based on swarm intelligence. A swarm of particles, i.e. candidate solutions, "fly" together to find the best solutions of the target problem. For particle i, $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$ represents the position and $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$ represents its velocity, where D represents the dimensionality of the search space. Each particle maintains its best position visited so far, i.e. personal best or *pbest*, and the best solution obtained by the whole swarm, i.e. global best or *gbest*. PSO updates x_i and v_i of each particle iteratively to find the optimal solutions.

PSO was firstly proposed to address continuous problems. Later, a binary PSO (BPSO) was developed in [15] to use PSO to address binary problems, where x_i, *pbest* and *gbest* can only have values 0 or 1. v_i shows the probability of x_i being 1. The v_i and x_i are updated using Equations (1), (2) and (3).

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_{i1} * (p_{id} - x_{id}^{t}) + c_2 * r_{i2} * (p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id} = \begin{cases} 1, \text{ if } rand() < s(v_{id}) \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

where

$$s(v_{id}) = \frac{1}{1 + e^{-v_{id}}} \quad (3)$$

where v_{id}^{t+1} is the velocity value of particle i in dth dimension in the $(t+1)$th iteration. w is the inertia weight, indicating influence of previous velocity. c_1 and c_2 are acceleration constants. r_{i1}, r_{i2} and $rand()$ are random values from [0, 1]. p_{id} and p_{gd} shows the values of $pbest$ and $gbest$. $s(v_{id})$, the sigmoid function, is to scale the velocity to (0, 1). A predefined maximum velocity, v_{max}, is to limit v_{id}^{t+1} to $[-v_{max}, v_{max}]$. $rand()$ is randomly chosen from [0,1].

B. Related Work on Dimension Reduction

This work focuses mainly on wrapper dimension reduction in classification. This section reviews related methods, including non-EC methods and EC based methods. Since this paper involves mainly on EC for dimension reduction, more space is given to EC based methods and only typical non-EC based methods are reviewed here, but readers are referred to recent surveys on (non-EC based) dimension reduction methods from [16], [6], [17].

1) Traditional Dimension Reduction Methods: Sequential forward selection (SFS) and sequential backward selection (SBS) [1] are two commonly used dimension reduction algorithms using the idea of greedy hill-climbing to search for the optimal attribute subset, but with different starting points, i.e. SFS starts with an empty set while SBS starts with the full set of attributes. However, both SFS and SBS may easily stuck into local optima [1] and they are expensive when the number of attributes is large [1]. Stearns [18] proposed a method named "plus-l-take away-r" where SFS was applied l times and then SBS was applied for r back tracking steps. However, the best values of (l, r) is challenging to determine.

Later, [19] proposed two floating selection methods, sequential backward floating selection (SBFS) and sequential forward floating selection (SFFS), which can automatically determine the (l, r) values. These two values are dynamically controlled, rather than being fixed values. Based on the best-first algorithm and SFFS, A linear forward selection (LFS) [20] was proposed, which restricted the number of attributes to be considered in each step. LFS further improves the efficiency of SFS while having comparable classification performance.

2) EC Technique to Dimension Reduction: There have been many EC-based dimension reduction methods, such as PSO based [21], [22], GP based [23], and differential evolution (DE) [24].

A combination of a GA and local search was employed to achieve dimension reduction in [25], where GA employed a

wrapper fitness function and local search employed a filter criterion. This memetic dimension reduction method achieved better performance than using GAs only and other methods.

A multi-tree GP method was used in [23] for dimension reduction, which was designed to select a attribute subset and simultaneously learn a classifier using the selected attributes. Kanan and Faez [26] proposed a wrapper attribute selection method using ant colony optimisation, which achieved better performance than other existing methods to solve a face detection problem. However, the performance of this method have not been tested on other benchmark problems. Wang et al. [27] developed DE based method for simultaneously dimension reduction and instance reduction, where a binary DE algorithm inspired by the idea of binary PSO was used to choose only informative attributes and representative instances from a datasets to reduce the size of the dataset. Experimental results showed that the proposed method could significantly reduce the data size and achieve comparable performance to the original data. A binary DE method was proposed for dimension reduction using a mutual information based filter evaluation [24].

Recently, BPSO has been applied to dimension reduction problems. Xue et al. [21] developed a PSO based two-stage dimension reduction algorithm to maximise the classification performance in the first phase and consider the number of attributes in the second phase. The experiments show that the two-stage algorithm can select a smaller number of attributes and achieve higher classification performance than other two PSO based dimension reduction algorithms. Xue et al. [28] used PSO for dimension reduction and proposed initialisation strategies and $pbest$ and $gbest$ updating strategies. Experiments indicated that the ideas could increase the accuracy and decrease the size of attribute subset and the computational cost. PSO was used for dimension reduction and clustering in machine learning [29], where each particle aimed to find optimal weights for all attributes and group the center values. Dimension reduction is accomplished by removing attributes with low weight values. However, because of interactions between attributes, attributes with low weights may be useful as well, but the removal of such attributes may decrease the performance of the attribute subset.

Statistical approaches have also been used to reduce the dimensionality of a dataset, e.g. Relief [30] uses a statistical metric to choose relevant attributes, in which each attribute has a score value showing its relevance to the class labels. Relief selects all relevant attributes, which may lead to a attribute subset including redundancy because relevant attributes can be redundant to each other. Many other measures, such as Pearson's correlation or least square regression error [31], have also been used in dimension reduction to score the discrimination ability of attributes in class separation [6].

Statistical variable grouping analysis is an important topic in statistics which aim to group attributes (or variables) to a number of groups. A statistical variable grouping algorithm considers interactions between attributes and groups relatively homogeneous attributes together [9], [10]. So, the interaction information between attributes found by a statistical variable grouping method, which is shown in the attribute groups, could be utilised to design a good dimension reduction algorithm, but this has not been investigated (except for our initial work [12], [13]). Since PSO is a powerful search technique in dimension reduction problems, this work will investigate the use of statistical variable grouping information and PSO for dimension reduction.

III. The Proposed Algorithms

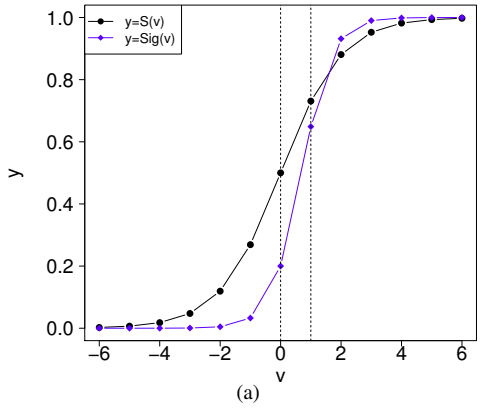
A recent statistical models based grouping method [9], [10] is used in this study to group attributes into different groups. The statistical attribute grouping method is conducted as a pre-processing step on a small subset of training instances to group attributes into different groups. By doing so, the PSO search will based on groups of attributes with the expectation of improving the efficiency of dimensionality reduction.

In this section, two new single objective algorithms are proposed to utlise the information discovered by the statistical variable grouping method, where the first algorithm aims to select a single attribute from each attribute group and the second algorithm allows the selection of multiple attributes from a single/each group. The two new methods select attributes based on *attribute groups*. They are different from the existing PSO based approaches, which select attributes based on the *whole attribute set*. To consider the variable grouping information during the search process, new position updating mechanisms are needed to develop the two new algorithms.

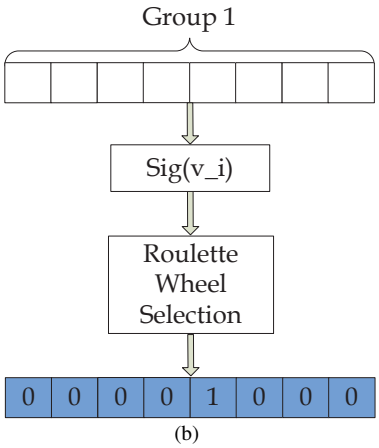
A. PSO with Roulette Wheel Dimension Reduction (PSORWS)

PSORWS is developed to maximise the classification accuracy by selecting a *single* attribute from each group. The rationale is that attributes from the same group are suppose to be similar and a single attribute can be chosen as the representative of the associate attribute group.

In BPSO for dimension reduction, a binary vector is used to encode a particle with the length or dimensionality as the total number of attributes in the dataset and each dimension corresponds to one attribute. For a particle, "1" in means the corresponding attribute is selected and "0" otherwise. The velocity value is transformed by Equation 3, $S(v)$, to the range of (0,1), which indicates the probability of the position value updates to "1" and also the probability of the corresponding attribute being selected. Based on this, a maximum probability based algorithm was proposed in [12], which always selects the attribute with the largest probability from each group. However, one potential limitation is that the swarm will easily lose the diversity and stagnation in local optima. Based on further investigation, it is found that the attribute with the largest $S(v)$ value in Equation 3 should have the largest probability to be selected, but the attributes with small velocities should also have (small) chances to be selected. To address this issue, we introduce a roulette wheel selection to select one attribute from each group based on the probability values. It is worth to mention that roulette wheel



(a)



(b)

Fig. 1. (a) Plot of Equation 3, $S(v)$ and Equation 4, $Sig(v)$; (b) Position update in PSORWS.

selection is based on the attributes' probability values within a group (not on individuals within the population).

The gradient of the original sigmoid function (Equation 3) is found not to be steep enough, which causes a small difference in probability between attributes that are significantly differing in velocity. Therefore, a new sigmoid function (Equation 4) is introduced to transform the original velocity to (0,1) instead of Equation 3. Fig. 1(a) shows the difference between Equation 3 and Equation 4, where the increase of v in the middle range will cause a greater increase in $Sig(v)$ than in $S(v)$. Therefore, Equation 4 better suits the roulette wheel selection approach than Equation 3 because it provides a greater bias towards the attributes with higher velocities. After applying Equation 4, the roulette wheel selection is performed to select one attribute from each group.

$$Sig(v_i) = \frac{1}{1 + 4e^{-2v_i}} \tag{4}$$

Fig. 1(b) uses one group/group as an example to show the position updating of PSORWS. Algorithm 1 describes the pseudo-code of PSORWS.

Algorithm 1: Pseudo-code of PSORWS

```
begin
    initialise x and v of all particles;
    randomly select one attribute from each group;
    while Maximum Iterations has been not reached do
        calculate classification performance of the selected
          attributes;
        for i=1 to SS(Swarm Size) do
            update pbest and gbest of particle i;
        for i=1 to SS do
            for d=1 to D do
                calculate v_i using Equation 1 ;   /* Update
                  velocity */
                calculate Sig(v_id) using Equation 4;
            for c=1 to Clusters Size do
                perform roulette wheel selection;
                update the dimension selected by roulette
                  wheel selection to 1 ;           /* Update
                  position */
                set the other dimensions for attributes in the
                  same groups to 0 ;               /* Update
                  position */
    calculate classification performance of the selected
      attributes on both the training and test sets;
    return gbest, the training performance and testing
      performance.
```

B. Improved Gaussian Based PSO for Dimension Reduction

PSORWS aims to select one attribute from each group, which may limit the classification performance because attributes from the same group might still contain complimentary information [32]. Therefore, allowing selecting *multiple* attributes from one group may increase the classification performance. In [13], we proposed such an algorithm (named GPSO) based on a Gaussian based position updating mechanism. The Gaussian updating mechanism is based on a Gaussian distribution function to consider the grouping information. It firstly determines the number of attributes to be selected from a certain group, and then select specific individual attributes from that group. The details can be seen from [13]. In this paper, we will further improve GPSO by proposing a new Gaussian based fitness function to develop an improved Gaussian based PSO algorithm (named IGPSO).

The new Gaussian based fitness function is developed to consider classification performance and the number of attributes. The classification performance can be easily shown by the accuracy of the selected attributes. The number of selected attributes cannot be directly used in the fitness function. The main reasons are that the number of attributes is not in the same range as the accuracy and that the number of selected attributes itself cannot reflect the grouping information. Therefore, we develop a new Gaussian criterion representing the number of attributes.

a) *Gaussian based similarity measure:* x represents the position of a particle (indicating a attribute subset), which is n-dimensional vector with n as the total number of attributes in the dataset. $\hat{x}$ is a C-dimensional vector, where C is the

number of attribute groups. Each dimension in $\hat{x}$ shows the number of attributes selected by x from the corresponding group. $\hat{y}$ is also a C-dimensional vector with each dimension value as "1", which means selecting 1 attribute from each group. $\hat{y}$ shows an ideal solution for the number of attributes when using the variable grouping information. Therefore, we develop a new criterion based on multivariate Gaussian distribution for measuring the similarity between $\hat{x}$ and $\hat{y}$, which will be used in the new fitness function to reflect the number of attributes.

For a dataset with C attribute groups, a C-dimensional multivariate Gaussian distribution is built. Since $\hat{y}$ contains 1 attribute from each group, a mean of 1 is chosen on each dimension of the multivariate Gaussian distribution. The logarithmic function, $\sigma = log(10 \times |clu|)$ developed earlier [13], is also used here to calculate the σ in each dimension, which ensures the variance along each dimension scales with the size of the attribute group. Since the Gaussian distribution is multivariate, a $C \times C$ covariance matrix is needed to establish the variances in each dimension. To achieve this, σ in each dimension is squared as to equal the variances and are placed on the diagonal entries of a $C \times C$ covariance matrix. By assigning all non-diagonal entries values as 0, it ensures that the selection of attributes from one group is independent from other groups.

Based on the covariance matrix, a Gaussian based similarity measure is developed, which is shown by Equation (5).

$$Sim(\hat{x}, \hat{y}) = \frac{\exp(-\frac{1}{2}(\hat{x} - \hat{y})^T Cov^{-1}(\hat{x} - \hat{y}))}{2\pi^{\frac{|clu|}{2}} \sqrt{|Cov|}} \quad (5)$$

where Cov is the multidimensional covariance matrix for the attribute group, Cov^{-1} is its inverse and $|Cov|$ is its determinant. $0 < Sim(\hat{x}, \hat{y}) \leq 1$. $Sim(\hat{x}, \hat{y}) = 1$ is the optimal/maximised value, which means $\hat{x}$ is the same as $\hat{y}$, i.e. selecting one attribute from each group. A solution that contains many attributes per group will have a Gaussian similarity measure score close to 0, which is the worst case.

b) Fitness function: Based on the proposed Gaussian based measure, a new fitness function is proposed, which is shown by Equation (6).

$$Fitness(x) = \alpha * Accuracy + (1 - \alpha) * Sim(\hat{x}, \hat{y}) \quad (6)$$

where "$Accuracy$" shows the classification accuracy of attributes selected by x. $Sim(\hat{x}, \hat{y})$ evaluates how close the $\hat{x}$ to $\hat{y}$, which indirectly reflects the number of attributes. α is set to [0,1] to balance the accuracy and the number of attributes. Obviously, Equation (6) is a maximisation function.

Based on the Gaussian updating mechanism developed earlier [13], the dimension reduction algorithm named GPSO was proposed. Based on the Gaussian fitness function, we further develop an improved Gaussian based PSO for dimension reduction algorithm (named IGPSO). The main difference between IGPSO and GPSO [13] is that IGPSO uses the newly developed fitness function while GPSO used the fitness function including the classification accuracy only. Comparing IGPSO with GPSO can discover whether the Gaussian fitness function can further reduce the number of attributes.

TABLE I
DATASETS

Dataset	#attributes	#**groups**	#classes	#instances
Wine	13	6	3	178
Australian Credit	14	7	2	690
Vehicle	18	6	4	846
German	24	10	2	1000
WBCD	30	6	2	569
Ionosphere	34	11	2	351
Lung	56	5	3	32
Sonar	60	12	2	208
Musk1	166	14	2	476
Arrhythmia	279	15	16	452

IV. Experimental Design

To test the performance of the proposed attribute selection algorithms, four benchmark methods are used as baseline algorithms, including two widely used conventional algorithms (i.e. *LFS* proposed in [20] and greedy stepwise backward selection (*GSBS*) from [33]), the random forest classification algorithm which includes attribute selection, a standard BPSO method (*PSOFS*) [28] without considering statistical variable grouping information, a statistical variable grouping based algorithm [12] (selecting one attribute per group), and a Gaussian based algorithm [13]. To save space, the detailed results of the methods from [13] and [12] are not presented in the next section, but the comparisons will be discussed.

Table I shows ten commonly used datasets from the UCI machine learning repository [34], including different numbers of attributes, classes and instances. Each dataset is randomly split into a training set (including 70% of the instances) and a test set (30% of the instances) [6], [3]. The training instances are used in the statistical variable grouping method to group attributes into different attribute groups. The number of groups obtained are also listed in Table I. Experiments have been conducted on ten different training and test partitions on each dataset using different random seeds. Similar patterns have been observed, so only the results on one partition are presented in the main paper and the results on the other nine partitions are shown in the appendices.

As wrapper approaches, a classification/learning algorithm is needed to evaluate the classification accuracy of the attribute subset. Any classification algorithm can be used here and a simple and commonly used algorithm, K-Nearest Neighbour (KNN) with K=5, is used here. The parameters of the PSO based algorithms are set as follows [14], [35]: w = 0.7298, c_1 = c_2 = 1.49618, v_{max} = 6.0, the population size is 30 with maximum 100 iterations. The fully connected topology is employed. α in the fitness function (Equation 6) is set as 0.98 to make sure the classification accuracy is much more important than the number of attributes. Experiments of LFS,

GSBS and random forest algorithms are run using Weka [36] and all the settings are kept to the defaults.

On each dataset, each algorithm was run for 40 times with different seeds. The Wilcoxon test non-parametric statistical significance test is applied on the testing accuracies to compare different methods, and the confidence interval is 95%.

V. Results and Discussions

Table II shows the experimental results of the new single objective methods PSORWS and IGPSO as well as PSOFS and Table III shows the experimental results of the two traditional methods (LFS and GSBS).

A. PSORWS: Results and Comparisons

In Table II, "All" means accuracy achieved by using all attributes for classification. Since each PSO method has been run for 40 independent runs, "AveSize" represents the average number of selected attributes, "BestAcc", "AveAcc" and "StdAcc" are the highest, mean and the standard deviation of the testing accuracies, respectively. "Test" represents the results of the significance test, where "+", "-" or "=" shows that PSOFS, PSORWS or IGPSO is significantly better, significantly worse, or similar to "All'.

As can be seen from Table II, the PSOFS method selected a small subset of attributes with less than half of the attributes, but achieved similar or significantly better classification performance than using all the attributes. The results suggest that PSOFS can be effectively used for dimension reduction.

According to Table II, the PSORWS algorithm selected attribute subsets with significantly higher or similar accuracy than using all attributes on most datasets. Meanwhile, since PSORWS was designed to select one attribute from each group, the number of attributes selected by PSORWS is significantly smaller than the total number of attributes.

Comparing PSORWS with PSOFS, it can be seen that the overall classification performance of PSORWS on the twelve datasets are similar to that of PSOFS. However, PSORWS selected much smaller attribute sets than PSOFS, especially on the datasets with a relatively large number (more than 50) of attributes, where dimension reduction is more necessary and more important than on datasets with a smaller number of attributes.

Comparing with our previous work [12], which is also based on statistical variable grouping information to select one attribute from each group, PSORWS achieved better classification performance than the algorithms developed in [12]. The main reason is that PSORWS introduces a greater amount of stochasticity, which provides a better search strategy that maintains greater swarm diversity helping the algorithm avoid stagnation in local optima. This allows the PSORWS algorithm to further improve the classification accuracy over the algorithms in [12] on most datasets, especially those having a larger number of attributes, which often have a more complex solution space.

Comparing PSORWS with LFS and GSBS in Table III, in most cases, although selecting a slightly larger number of attributes, PSORWS achieved better classification performance than LFS. PSORWS outperformed GSBS in terms of the classification accuracy and the number of attributes on all the ten datasets.

TABLE II
Results

Dataset	Algorithm	AveSize	BestAcc	AveAcc ± StdAcc	Test
Wine	All	13	76.54		
	PSOFS	8.32	97.53	95.96 ± 1.87E0	+
	PSORWS	6	100	100 ± 0E0	+
	IGPSO	5.65	98.77	96.91 ± 2.83E0	+
Australian	All	14	70.05		
	PSOFS	4	87.44	87.44 ± 4.26E-14	+
	PSORWS	7	79.23	79.23 ± 5.68E-14	+
	IGPSO	3.02	87.44	85.56 ± 30.2E-2	+
Vehicle	All	18	83.86		
	PSOFS	9.28	85.83	84.3 ± 61.9E-2	+
	PSORWS	6	83.66	83.47 ± 3.07E-2	-
	IGPSO	8.55	84.65	83.87 ± 39.4E-2	=
German	All	24	68.0		
	PSOFS	12.9	72	68.73 ± 1.3E0	+
	PSORWS	10	71	70.83 ± 50E-2	+
	IGPSO	9.6	73.33	70.12 ± 1.57E0	+
WBCD	All	30	92.98		
	PSOFS	14.92	92.98	92.98 ± 0E0	=
	PSORWS	6	94.74	94.43 ± 34.6E-2	+
	IGPSO	6.92	94.74	93.11 ± 38.1E-2	+
Ionosphere	All	34	83.81		
	PSOFS	10.38	93.33	89.05 ± 1.84E0	+
	PSORWS	11	92.38	88.71 ± 65.8E-2	+
	IGPSO	8.65	92.38	89.79 ± 1.55E0	+
Lung	All	56	70.0		
	PSOFS	26.92	80	72.5 ± 5.36E0	+
	PSORWS	5	90	89.75 ± 1.56E0	+
	IGPSO	7.8	90	83 ± 7.14E0	+
Sonar	All	60	76.19		
	PSOFS	24.72	87.3	79.52 ± 2.92E0	+
	PSORWS	12	80.95	75.91 ± 2.51E0	=
	IGPSO	16.32	82.54	77.42 ± 3.18E0	+
Musk1	All	166	83.92		
	PSOFS	83.6	89.51	85.65 ± 2.1E0	+
	PSORWS	14	86.71	81.15 ± 2.57E0	-
	IGPSO	36.05	88.81	84.56 ± 2.35E0	=
Arrhythmia	All	279	94.46		
	PSOFS	119.35	95.14	94.57 ± 33.5E-2	=
	PSORWS	15	95.59	94.75 ± 33.1E-2	+
	IGPSO	27.2	95.25	94.49 ± 40.8E-2	=

The results show that the PSORWS algorithm can successfully utilise the statistical variable grouping information to address dimension reduction problems by selecting a single attribute from each group, so that it is able to reduce the dimensionality without decreasing but often increasing the accuracy. However, although attributes from the same group are similar attributes, they still can be complimentary to each other (particularly for large groups) and need to be used together to benefit the classification process due to interactions between attributes. Therefore, PSORWS has a potential limitation of missing complimentary attributes because of selecting only one attribute from each group.

B. IGPSO: Results and Comparisons

According to Table II, it can be observed that on average, IGPSO selected a *much* smaller attribute subset than the original attribute set in *all* cases. By using the selected

TABLE III
EXPERIMENTAL RESULTS OF LFS AND GSBS

Method	Wine		Australian		Vehicle		German		WBCD	
	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy
LFS	7	74.07	4	70.05	9	83.07	3	68.67	10	88.89
GSBS	8	85.19	12	69.57	16	75.79	18	64.33	25	83.63
Method	Ionosphere		Lung		Sonar		Musk1		Arrhythmia	
	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy	Size	Accuracy
LFS	4	86.67	6	90.0	3	77.78	10	85.31	11	94.46
GSBS	30	78.1	33	90.0	48	68.25	122	76.22	130	93.55

attributes, the classification performance of IGPSO is similar or significantly better than using all attributes on *all* datasets. On the three datasets that have more than 200 attributes (i.e. the Arrhythmia dataset), IGPSO selected less than 10% of the original attributes and achieved significantly higher accuracy than using all attributes on two datasets, and achieved similar results on the other datasets.

Compared with PSOFS, IGPSO achieved similar or better performance than the PSOFS algorithm. Meanwhile, the average number of attributes obtained by IGPSO is always smaller or even much smaller than the PSOFS method on *all* datasets.

Compared with PSORWS, the accuracy achieved by IGPSO is at least similar than PSORWS in almost all cases. The number of attributes is slightly larger in some cases (which is expected). This suggests that by allowing selecting multiple attributes from each group, IGPSO can successfully improve the classification performance, but not always increase the number of attributes (such as in Wine and Australian datasets) since it allows selection of zero attributes from some groups if all attributes in those groups are "irrelevant".

In our previous work [13], a PSO based algorithm named GPSO was proposed to select multiple attributes from each group. The main difference between GPSO and IGPSO is that IGPSO uses the newly developed fitness function, which considers the number of attributes selected from each group. Comparing Table II with the results in [13], IGPSO further reduced the number of attributes in almost all cases, especially on high dimensional datasets. The main reason for this lies on the Guassian based measure, which reflects the number of attributes in the new fitness function in IGPSO, plays a significant role on datasets with a larger number of attributes. In terms of the classification accuracy, IGPSO is similar to GPSO in most cases, but a significant reduction of the number of attributes in IGPSO results in a slight reduction in its classification performance, such as on Arrhythmia. IGPSO in general is a better approach than GPSO to dimension reduction, because IGPSO achieved similar accuracy to GPSO while selected a much smaller attribute subset than GPSO on higher dimensional datasets, where dimension reduction is more important and necessary.

Comparing IGPSO with LFS and GSBS in Table III, it can be seen that in all cases, IGPSO achieved higher classification accuracy than both LFS and GSBS. IGPSO selected a larger number of attributes than LFS, but a significantly smaller number of attributes than GSBS in all cases. Since in most cases, classification performance is more important than the number of attributes, IGPSO is a better dimension reduction method than LFS and GSBS. PSORWS outperformed GSBS in all cases, i.e. higher accuracy and lower number of attributes.

The results suggest that by using the Gaussian based updating mechanism and the Gaussian fitness function, IGPSO can successfully use the statistical variable grouping information to address dimension reduction problems. IGPSO reduced the dimensionality of the datasets and at the same time increased the classification performance in *all* cases. IGPSO also outperformed PSOFS, PSORWS and the previously developed algorithm [13] without using the new Gaussian fitness function.

VI. CONCLUSIONS AND FUTURE WORK

The goal of this paper was to propose a new dimension reduction approach in classification based on PSO and statistical variable grouping. The goal has been successfully achieved by developing two new wrapper methods, where the first algorithm PSORWS which selects a single attribute from each group to reduce the number of attributes, and the second algorithm named IGPSO that utlises the grouping information, allowing the selection of multiple (or zero) attributes from one group. The proposed algorithms were compared with LFS and GSBS, a PSO algorithm without using statistical variable grouping (PSOFS), and a PSO algorithm without using statistical variable grouping (IGPSO) on twelve datasets of varying difficulties.

Experiments show PSORWS and IGPSO can successfully use the statistical variable grouping information to reduce the number of attributes, and keep the same or improve the performance of classification. Specifically, PSORWS significantly reduced the dimensionality and improved the classification performance. IGPSO further increased the classification performance with a slight increase in the number of attributes. Both of the new algorithms outperformed the existing conventional and PSO based algorithms.

This work demonstrates the benefits of using statistical variable grouping and the Gaussian fitness function. In future work, we will utilise such information in multi-objective attribute selection, which requires a multi-objective search mechanism to simultaneously maximising the classification accuracy and minimising the number of attributes. We will also develop novel filter approaches, which are more general and also computationally cheaper than wrappers. Furthermore,

the analysis of interactions between attributess discovered by the statistical variable grouping method could lead to further improvements in the classification performance of the selected attributes. However, this will need domain knowledge. Since the proposed algorithms are wrapper approaches, the computational time on large datasets can be very long. We also intend to develop a new approach combining filter and wrapper methods to reduce computational cost without increasing (or even further decreasing) the classification error rate and the number of attributes.

REFERENCES

[1] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[2] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2016.

[3] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 29, pp. 29– 38, 2008.

[4] Y. Mei, B. Xue, and M. Zhang, "Fast bi-objective feature selection using entropy measures and bayesian inference," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '16. New York, NY, USA: ACM, 2016, pp. 469–476.

[5] I. Babaoglu, O. Findik, and E. Ulker, "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3177 – 3183, 2010.

[6] B. Xue, M. Zhang, W. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[7] A. P. Engelbrecht, *Computational intelligence: an introduction (2. ed.)*. Wiley, 2007.

[8] H. B. Nguyen, B. Xue, and P. Andreae, "Mutual information for feature selection: estimation or counting?" *Evolutionary Intelligence*, vol. 9, no. 3, pp. 95–110, 2016.

[9] S. Pledger and R. Arnold, "Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection," *Computational Statistics & Data Analysis*, vol. 71, no. 0, pp. 241–261, 2014.

[10] E. Matechou, I. Liu, S. Pledger, and R. Arnold, "Biclustering models for ordinal data," 2011, presentation at the NZ Statistical Assn. Annual Conference, University of Auckland.

[11] E. Alpaydin, *Introduction to machine learning*. The MIT Press, 2004.

[12] M. Lane, B. Xue, I. Liu, and M. Zhang, "Particle swarm optimisation and statistical clustering for feature selection," in *AI 2013: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, 2013, vol. 8272, pp. 214–220.

[13] ——, "Gaussian based particle swarm optimisation and statistical clustering for feature selection," in *14th European Conference on Evolutionary Computation in Combinatorial Optimization (EvoCOP)*, ser. Lecture Notes in Computer Science, 2014, pp. 133–144.

[14] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *IEEE International Conference on Evolutionary Computation (CEC'98)*, 1998, pp. 69–73.

[15] J. Kennedy and R. Eberhart, "A discrete binary version of the particle swarm algorithm," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, 1997, pp. 4104–4108.

[16] Y. Zhai, Y.-S. Ong, and I. Tsang, "The emerging "big dimensionality"," *IEEE Computational Intelligence Magazine*, vol. 9, no. 3, pp. 14–26, 2014.

[17] B. Xue, M. Zhang, and W. N. Browne, "A comprehensive comparison on evolutionary feature selection approaches to classification," *International Journal of Computational Intelligence and Applications*, vol. 14, no. 02, p. 1550008, 2015.

[18] S. Stearns, "On selecting features for pattern classifier," in *Proceedings of the 3rd International Conference on Pattern Recognition*. IEEE Press, 1976, pp. 71–75.

[19] P. Pudil, J. Novovicova, and J. V. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[20] M. Gutlein, E. Frank, M. Hall, and A. Karwath, "Large-scale attribute selection using wrappers," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM '09)*. IEEE, 2009, pp. 332–339.

[21] B. Xue, M. Zhang, and W. N. Browne, "New fitness functions in binary particle swarm optimisation for feature selection," in *IEEE Congress on Evolutionary Computation (CEC'12)*, 2012, pp. 2145–2152.

[22] B. H. Nguyen, B. Xue, and P. Andreae, "A novel binary particle swarm optimization algorithm and its applications on knapsack and feature selection problems," *Intelligent and Evolutionary Systems*, pp. 319 – 332, 2017.

[23] D. Muni, N. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, pp. 106–117, 2006.

[24] X. He, Q. Zhang, N. Sun, and Y. Dong, "Feature selection with discrete binary differential evolution," in *International Conference on Artificial Intelligence and Computational Intelligence (AICI '09)*, vol. 4, 2009, pp. 327–330.

[25] Z. X. Zhu, Y. S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 1, pp. 70–76, 2007.

[26] H. R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (aco) evaluated on face recognition system," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 716–725, 2008.

[27] J. Wang, B. Xue, X. Gao, and M. Zhang, *A Differential Evolution Approach to Feature Selection and Instance Selection*. Springer International Publishing, 2016, pp. 588–602.

[28] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied Soft Computing*, vol. 18, no. 0, pp. 261–276, 2014.

[29] M. Javani, K. Faez, and D. Aghlmandi, "Clustering and feature selection via pso algorithm," in *International Symposium on Artificial Intelligence and Signal Processing (AISP'11)*, 2011, pp. 71 –76.

[30] K. Kira and L. A. Rendell, "A practical approach to feature selection," *Assorted Conferences and Workshops*, pp. 249–256, 1992.

[31] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 2, pp. 1330–1339, 2009.

[32] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[33] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.

[34] K. Bache and M. Lichman, "Uci machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[35] M. Clerc and J. Kennedy, "The particle swarm– explosion, stability, and convergence in a multidimensional complex space," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58–73, 2002.

[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, pp. 931–934, 2009.