Binary Particle Swarm Optimisation and Rough Set Theory for Dimension Reduction in Classification

Liam Cervante[1], Bing Xue[1], Lin Shang[2], Mengjie Zhang[1]
[1] School of Engineering and Computer Science
Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
[2]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210046, China
Email: {Liam.Cervante, Bing.Xue, Mengjie.Zhang} @ecs.vuw.ac.nz, shanglin@nju.edu.cn

***Abstract*—Dimension reduction plays an important role in many classification tasks. In this work, we propose a new filter dimension reduction algorithm (PSOPRSE) using binary particle swarm optimisation and probabilistic rough set theory. PSOPRSE aims to maximise a classification performance measure and minimise a newly developed measure reflecting the number of attributes. Both measures are formed by probabilistic rough set theory. PSOPRSE is compared with two existing PSO based algorithms and two traditional filter dimension reduction algorithms on six discrete datasets of varying difficulty. Five continues datasets including a large number of attributes are discretised and used to further examine the performance of PSOPRSE. Three learning algorithms, namely decision trees, nearest neighbour algorithms and naive Bayes, are used in the experiments to examine the generality of PSOPRSE. The results show that PSOPRSE can significantly decrease the number of attributes and maintain or improve the classification performance over using all attributes. In most cases, PSOPRSE outperforms the first PSO based algorithm and achieves better or much better classification performance than the second PSO based algorithm and the two traditional methods, although the number of attributes is slightly large in some cases. The results also show that PSOPRSE is general to the three different classification algorithms.**

I. Introduction

Classification tasks usually include a large number of attributes and suffer from "the curse of dimensionality", which refers to the decrease of a known classification method's performance caused by the increase of the number of attributes [1, 2]. To solve this problem, dimension reduction is introduced to remove unnecessary attributes to reduce the dimensionality while preserving the representation power of the original high-dimensional attributes to maintain the classification performance [2]. By removing the unnecessary attributes, dimension reduction can reduce the training time of a learning algorithm, facilitate data visualization, simplify the learnt classifier, and/or increase the classification performance [2, 3].

Dimension reduction is a challenging task due mainly to two reasons, which are attribute interaction and the large search space. Attribute interaction may lead to the phenomenon that attributes, which are individually relevant (irrelevant or redundant or) to class labels, may become redundant (relevant) if they are combined together with other attributes for classification. The best group of attributes should be complementary to each other. The evaluation criterion, which measures the goodness of the selected attributes, is an important issue in dimension reduction. Based on the evaluation criteria, dimension reduction methods are generally classified into two broad classes: wrapper approaches and filter approaches [2]. Wrapper approaches include a learning/classification method to evaluate the selected attributes. Therefore, wrappers often obtain better classification performance than filter approaches, but they suffer from the high computation cost and the loss of generality, i.e. specific to a particular classification algorithm. Filter approaches are independent of any learning algorithm. Therefore, filter approaches usually need a good evaluation criterion.

The size of search space in dimension reduction problems increases exponentially along with the number of attributes. An exhaustive search is impractical in most situations. Different heuristic search techniques have been applied to dimension reduction problems [1], but most of them still have the limitations of long computation time and being stuck in local optima [2]. Therefore, a computationally cheap global search algorithm is needed to better solve dimension reduction problems. Evolutionary computation (EC) techniques are well-known global search algorithms, which have been used for dimension reduction, including genetic programming (GP) [4], genetic algorithms (GAs) [5], and particle swarm optimisation (PSO) [3]. PSO is a relatively recent EC algorithm and is computationally less expensive than GAs and GP. PSO has been successfully applied to dimension reduction [3, 6, 7].

Most existing EC based dimension reduction algorithms are wrapper approaches. The use of such algorithms is limited in real-world applications due to the long computation time. The development of EC based filter dimension reduction approaches is still an open issue. Rough set theory [8] is able to deal with uncertainty, imprecision and vagueness, which has been successfully used for dimension reduction [9]. However, standard rough set theory has some limitations [10]. Therefore, Yao and Zhao [10] developed probabilistic rough set theory, but this idea has not been implemented for dimension reduction by other researchers. A filter dimension reduction algorithm using PSO and probabilistic rough set was developed in [6] and obtained better performance than using PSO and standard rough set. However, the proposed algorithm in [6] needs to define a parameter to balance the relative importance assigned for the classification performance and the number of attributes, which is problem-dependent and difficult to determine in advance. Meanwhile, due to the the constraint that rough set theory only works on discrete data, the datasets used in rough set in recent work [9, 11, 6, 12] only have a small number of attributes.

A. Goals

The overall goal of this research is to develop a filter dimension reduction approach using PSO and probabilistic rough set theory to reduce the number of attributes and achieve similar or even better classification performance than using all attributes, which is not only for datasets with a small number of attributes, but also for datasets with a large number of attributes. To achieve this goal, a new fitness function formed by probabilistic rough set theory is proposed to maximise the representation/classification power and minimise the number of attributes, but does not need to predefine a parameter to balance the relative importance of these two components. The proposed algorithm is examined and compared with two existing PSO based algorithms and two traditional dimension reduction algorithms on six commonly used discrete datasets and five continues datasets with a large number of attributes. Specifically, we will investigate

- whether the newly proposed algorithm can reduce the number of attributes and maintain or increase the classification performance over using all the attributes,
- whether the proposed algorithm outperform two existing PSO based algorithms and two traditional dimension reduction algorithms,
- whether the proposed filter algorithm is general to different classification algorithms, and
- whether the proposed algorithm and the two existing PSO based algorithms can be used for datasets including a larger number of attributes.

II. Background

A. Binay Particle Swarm Optimisation (BPSO)

Particle swarm optimisation (PSO) [13, 14] is inspired by social behaviours, such as fish schooling and birds flocking. In PSO, each solution of the target problem is represented by a particle. A swarm of particles move ("fly") together in the search space to find the best solutions. For any particle i, a vector $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$ is used to represent its position and a vector $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$ is used to represent its velocity, where D means the dimension of the target problem. During the evolutionary process, each particle can remember its best position visited so far called personal best (denoted by *pbest*), and the best previous position visited so far by the whole swarm called global best (denoted by *gbest*). Based on personal best and global best, PSO iteratively updates x_i and v_i of each particle to search for the optimal solutions.

Originally, PSO was proposed to address problems/tasks with a continuous search space. To extend PSO to address discrete problems, a binary PSO (BPSO) was developed in [15], where x_i, *pbest* and *gbest* are limited to 0 or 1. v_i in BPSO represents the probability of an element in the position updating to 1. BPSO updates v and x of each particle according to Formulae 1 and 2.

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

$$x_{id} = \begin{cases} 1, \text{ if } rand() < \frac{1}{1+e^{-v_{id}}} \\ 0, \text{ o}therwise \end{cases} \quad (2)$$

where v_{id}^{t+1} shows the velocity of particle i in the dth dimension in the $t+1$th iteration. w is the inertia weight, which indicates the influence of the previous velocity. c_1 and c_2 are acceleration constants. r_{i1}, r_{i2} and $rand()$ are random valuess, which are uniformly distributed in [0, 1]. p_{id} and p_{gd} shows the values of personal best and global best in the dth dimension. A predefined maximum velocity, v_{max}, is to limit v_{id}^{t+1} to $[-v_{max}, v_{max}]$.

B. Probabilistic Rough Set Theory

Rough set theory was developed by Pawlak [8] to deal with uncertainty and imprecision. One of its advantages is that rough set does not need prior knowledge or additional information about data.

In rough set theory, the data of a problem is organised in a table called decision table. In the decision table, one row shows an object and one column corresponds to attributes in the dataset. Here, the decision table is denoted as $I = (U, A)$, where U is the universe of objects in the dataset and A is the collection of attributes that describe the objects. $A = C \cup D$, where C is the decision attribute indicating the class to which each object belongs and D shows all the other attributes which are called conditional attributes.

Partitions are the knowledge base of rough set theory, which are obtained according to equivalence relation defined in the universe. For any $P \subseteq A$ and $X \subseteq U$, the equivalence relation is defined as $IND(P) = \{(x, y) \in U | \forall a \in P, a(x) = a(y)\}$. The equivalence class of $IND(P)$ is denoted as $[x]_P$, which means that with regards to P, $\forall y \in [x]_P$, (x, y) are indiscernible to each other. Based on the equivalence classes, rough set theory defines the lower approximation (denoted by $\underline{P}X$) and the upper approximation (denoted by $\overline{P}X$) of the set target X with regards to P [8], which are shown as follows:

$$\underline{P}X = \{x \in U | [x]_P \subseteq X\} \quad (3)$$

$$\overline{P}X = \{x \in U | [x]_P \cap X \neq \emptyset\} \quad (4)$$

Regarding the relationships between the target set X and an equivalence class, $\underline{P}X$ and $\overline{P}X$ in the standard rough set were defined in two extreme cases. The objects in $\underline{P}X$ can be definitely classified to the target set X. $\overline{P}X$ includes the objects, which probably or definitely belong to the target set X. A pair of $(\overline{P}X, \underline{P}X)$ is called a rough set.

In standard rough set theory, the degree of the overlap between the target set X and an equivalence class is not taken into account. This will limit the application of standard rough set theory on many problems. Probabilistic rough set theory was proposed to avoid this limitation, where the definitions of the lower and upper approximation can be relaxed [10]. In probabilistic rough set theory, $\mu_P[x]$ defines the probability of the equivalence class $[x]_P$ also included in the target set A, which can be seen in Equation 5. Equation 6 defines the lower approximation, where α can be adjusted to restrict or relax the lower approximation.

$$\mu_P[x] = \frac{|[x]_P \cap X|}{|[x]_P|} \tag{5}$$

$$\underline{apr}_P X = \{x|\mu_P[x] \geq \alpha\} \tag{6}$$

Note that probabilistic rough set is based essentially on the majority rule. If the majority of an object x's equivalent objects in $[x]_P$ are in the target set X, the object x is put in the lower approximation of the target set X. $\underline{apr}_P X = \underline{P}X$ when $\alpha = 1$. $\underline{apr}_P X$ loosens the boundaries of the rough set if $\alpha < 1$.

In rough set theory, a reduct (related to a subset of conditional attributes) is the essential part of a decision table. The approximation power of a reduct should be similar to that of A, which includes all the original attributes. There could be many different reducts in a rough set system and dimension reduction aims to obtain the smallest reduct.

C. Related Work on Dimension Reduction

In recent years, a number of dimension reduction methods have been developed [2, 16]. Typical algorithms are briefly reviewed in this section.

1) Traditional Dimension Reduction Methods: Based on greedy search, two sequential methods, i.e. sequential forward selection (SFS) [17] and sequential backward selection (SBS) [18], are developed, which are typical wrapper dimension reduction methods. The starting point of SFS is an empty set of attributes while SBS starts with all the available attributes. Candidate attributes are sequentially selected (removed from) the attribute set until the next selection (removal) decreases or does not improve the classification performance. SFS and SBS suffer from the problem of nesting effect. Stearns [19] proposed the "plus-l-take away-r" method to address this problem. The proposed algorithm conducts l times forward selection and then r times backward elimination. However, it is difficult to determine the best values of (l, r).

Hall [20] proposes a filter correlation based dimension reduction method (Cfs), which employs the correlation between attributes and class labels as the evaluation criterion. Almuallim and Dietterich [21] propose a filter algorithm which performs an exhaustive search of all possible combinations of attributes, and selects the smallest subset. However, performing an exhaustive search is computationally expensive. Relief [22] is a filter algorithm in which each attribute has a score showing its relevance to the class labels and all the relevant attributes are selected. However, the attributes selected by Relief may still have redundancy, because the redundancy between relevant attributes is not taken into account.

2) EC Algorithms for Dimension Reduction: In recent years, EC algorithms have been used to address dimension reduction problems, such as GAs, GP, ant colony optimisation (ACO) and PSO.

Zhu et al. [5] propose a dimension reduction method using a memetic algorithm that combines local search and a GA. Individual attributes are firstly ranked according to a filter measure and GA adds or deletes an attribute according to the ranking results. Experiments indicate that the proposed algorithm obtains better results than GA and other algorithms. The results also suggest that the performance and efficiency can be improved by setting a proper balance between genetic algorithm and local search. Based on GP, Kourosh and Zhang [4] propose a relevance measure named GPRM, which is used to evaluate and rank attribute subsets in binary classification problems. Dimension reduction can be achieved by using a top-ranked subset that has a smaller number of attributes for classification. Ming [11] proposes a filter dimension reduction algorithm using ACO and rough set, which starts with the core attributes evaluated by rough set theory. Experiments indicate that the proposed method outperforms a C4.5 based dimension reduction algorithm in terms of both the number of attributes and the classification accuracy. However, it is not compared with any other EC based dimension reduction approaches.

As an EC algorithm, PSO has gained attention for addressing dimension reduction problems. Based on BPSO, Iswandy and Koenig [23] develop a filter based dimension reduction algorithm. The proposed algorithm employs different weights to linearly combine three objectives, which are evaluated by three filter criteria, into a single fitness function. The results indicate that this algorithm outperforms other methods on several benchmark problems. Wang et al. [9] develop a filter dimension reduction method using an improved BPSO and rough set model. However, the classification performance of the reduct is only tested on one learning algorithm, the LEM2 algorithm, which originally is specific used for rough set theory and have some bias for the proposed rough set theory based algorithm. Meanwhile, only using one learning algorithm to evaluate the classification performance can not show the advantage that filter algorithms are more general. Marinakis et al. [24] propose a wrapper dimension reduction approach based on BPSO and KNN for a real-world medical diagnosis problem, which is called Pap-smear cell classification problem. The proposed algorithm can remove around half of the features and achieves good classification performance.

Based on a filter measure and PSO, a filter-wrapper dimension reduction algorithm is proposed in [25], with the goal of integrating their advantages. The filter measure is used to encode the position of each particle and the classification performance is used in the fitness function. Experiments show that the proposed method slightly outperforms a BPSO based filter method. However, it has not been compared with any wrapper algorithm, which usually can obtain higher classification performance than a filter algorithm. Lin and Chen [3] propose a wrapper dimension reduction algorithm (PSOLDA) based on PSO and a linear discrimination analysis algorithm (LDA), which aims to maximise the classification performance evaluated by LDA. Different parameters are tuned to obtain the best settings for PSOLDA. Experimental results show that PSOLDA outperforms LDA using all attributes, LDA with principal components analysis (PCA), and LDA with forward and backward selection in almost all cases. However, PSOLDA is sensitive to parameter setting and the datasets in the experiments have a small number of attributes.

A dynamic quantum-inspired PSO algorithm is developed for dimension reduction and at the same time for parameter optimisation in neural networks [26]. Compared with two other standard PSO and quantum information based algorithms, the proposed method is computationally cheaper and obtains better classification accuracy. Bae et al. [12] applied an intelligent dynamic swarm based BPSO for dimension reduction, where rough set theory is employed as the evaluation criterion in the fitness function. K-mean algorithm is used to help the proposed algorithm to handle continuous data. The results suggest that the proposed method can overcome the premature convergence problem and shorten the computation time. However, the number of attributes in the datasets is small.

In summary, PSO has been successfully applied to dimension reduction problems. However, most existing dimension reduction algorithms are wrapper approaches, which are less general than filter algorithms and computational inefficiency. Therefore, it is an open issue to use PSO to develop a filter dimension reduction approach.

III. Proposed Approach

In this section, two existing dimension reduction algorithms [6] using PSO and probabilistic rough set are briefly described. They are used to compare with the performance of the proposed algorithm. Then we use probabilistic rough set theory to develop a new measure to reduce the number of attributes based on which a new algorithm is proposed.

A. Existing Algorithms: PSORS and PRORSN

In classification problems, the information of the data can be represented using the decision table in rough set theory. Each instance can be regarded as an object. The class label in the dataset is the decision attribute D and other attributes/features are the conditional attributes C and $A = C \cup D$. Therefore, based on the equivalence relation defined by A, U (all the instances in the dataset) can be partitioned to different target set or classes, $U_1, U_2, U_3, ..., U_n$, where n is the number of classes. Dimension reduction is to remove some of the attributes so that the remaining set $P \subseteq A$ contains a small number of attributes and maintains the information described by all the original attributes. The goodness of P can be measured by how well P represent each class (U_i) in the dataset (U).

a) PSORS: Since standard rough set theory has some limitations because of the definitions of lower approximation and upper approximation [10], probabilistic rough set theory was used in [6] to propose a dimension reduction algorithm (PSORS), where PSO was used as the search technique. In PSORS, for the target set U_1 in U in probabilistic rough set theory, $\mu_P[x] = \frac{|[x]_P \cap U_1|}{|[x]_P|}$. $\mu_P[x]$ quantifies the proportion of instances in the equivalence class $[x]_P$ also included in U_1. $\underline{apr}_P U_1 = \{x|\mu_P[x] \geq \alpha\}$ defines the lower approximation of U_1 with regards to P, which is how well P describes the target set U_1. According to probabilistic rough set theory, $[x]_P$ does not have to be completely contained in U_1. α can be adjusted to restrict or relax the definition of $\underline{apr}_P U_1$. How well P describes the universe U can be calculated by summing how well P describes each target set, which is shown by Equation 7. Therefore, Equation 7 is used as the fitness function of PSORS, which essentially measures the number of instances that P correctly makes distinguishable from other instances in the datasets. $Fitness(P) = 1$ means that all the instances are correctly identified to the true classes.

$$Fit_1 = \frac{\sum_{i=1}^{n} |\underline{apr}_P U_i|}{|U|} \tag{7}$$

b) PSORSN: PSORS using probabilistic rough set theory can avoid the limitations of standard rough set, but PSORS does not consider the number of attributes. During the evolutionary process, if there are two or more reducts that have the same fitness value, PSORS does not prefer the smaller ones. Therefore, the number of attributes was added into the fitness function to form another algorithm (PSORSN) in [6], which aims to maximise the representation power of the attribute subset and minimise the number of attributes at the same time, as shown in Equation 8.

$$Fit_2 = \gamma * \frac{\sum_{i=1}^{n} |\underline{apr}_P U_i|}{|U|} + (1-\gamma)*(1 - \frac{\#attributes}{\#totalFeatures}) \tag{8}$$

where $\gamma \in (0, 1]$ is the importance of the representation power of the attribute subset while $(1 - \gamma)$ indicates the importance of the number of attributes. When $\gamma = 1.0$, PSORSN is the same as PSORS.

B. New Algorithm: PSORSE

PSORSN considers the number of attributes in the fitness function, which is a typical way to combine two objectives into one single fitness function. However, in the situation of using probabilistic rough set theory, it might not work well for some problems/datasets. The main reason is that a very small number of attributes can describe a very large number of equivalence classes, which attempt to extract patterns in the dataset. However, each equivalence class may have a very small number of instances. For example, 20 binary attributes can describe 1048576 (2^{20}) equivalence classes. There could be thousands of small equivalence classes, which only contain one or two instances. If there is another equivalence class, which has slightly more instances, this class will dominate others and the obtained reduct will only contain information that can identify this particular class. Therefore, without considering the size of the equivalence classes, Fit_2 may obtain a small reduct, but may loss generality and not perform well on unseen test data.

In order to address the problem, we use probabilistic rough set theory to develop a new measure to minimise the number of attributes in the reduct, which aims to minimise the number of equivalence classes and maximise the number of instances

Algorithm 1: Pseudo-code of PSORSE, PSORS and PSORSN

```
begin
    split the instances into a Training and a Test set;
    initialise $x$ and $v$ of each particle;
    while $Maximum\ Iterations\ has\ been\ not\ met$ do
        calculate the fitness value each particle on the Training
        set according to Equation 7 for PSORS, Equation 8 for
        PSORSN and Equation 9 for PSORS;
        for i=1 to $Swarm\ Size$ do
            update the personal best ($pbest$) of particle $i$;
            update the global best ($gbest$) of particle $i$;
        for i=1 to $Swarm\ Size$ do
            for d=1 to $Dimensionality$ do
                calculate $v_i$ according to Equation 1
                calculate $x_i$ according to Equation 2
    calculate the classification performance of the selected
    attributes on the test set using DT, NB or 5NN as the
    classification algorithm;
    return the position of $gbest$ (the selected attributes);
    return the training and testing classification performance;
```

in each equivalence class. Based on this new measure, we propose a new PSO based dimension reduction algorithm (PSORSE), where Equation 9 is used as the fitness function.

$$Fit_3 = \frac{\sum_{x=1}^{n} |\underline{apr}_P X_i|}{|U|} + \frac{\sum_{x\in\{equivalence\ classes\}} \frac{|x|}{|\mathbf{U}|}}{\#\ of\ equivalence\ classes} \quad (9)$$

The pseudo-code of PSORSE, PSORS and PSORSN is shown in Algorithm 1. In all the three algorithms, each particle is represented by a binary string, whose length is the total number of attributes in the dataset, which also represents the dimension of the solution space. “0” in the binary string indicates that the corresponding attribute is removed and “1” indicates that this attribute is not removed.

IV. Design of Experiments

To examine the performance of the new approach, a set of experiments have been conducted on six datasets (listed in Table I), which are chosen from UCI machine learning repository [27]. These six datasets have different numbers of instances, attributes, and classes. They are used as representative examples of the tasks that the proposed method will test on. As rough set theory only works on discrete values, all the six datasets are categorical data. In each dataset, two thirds of the instances are chosen as the training set while others are used as the test set. The filter algorithms first run on the training set in order to select a set of attributes. The training process is independent of any classification algorithm. The performance of the selected attributes is then evaluated by a learning/classification algorithm on the unseen test set. Almost all learning algorithms that are able to deal with discrete data can be used here. Three different learning algorithms, decision trees (DT), naive Bayes (NB) and K-nearest neighbor algorithms with K=5 (5NN), are used in the experiments to test the claim that filter dimension reduction methods are general.

TABLE I
Datasets

Dataset	# Attributes	# Classes	# Instances
Lymphography (Lymph)	18	4	148
Dermatology	33	6	366
Soybean Large	35	19	307
Chess	36	2	3196
Waveform	40	3	5000
Statlog	36	6	6435

All the α values should be larger than 0.5 because the lower approximation in probabilistic rough set theory defines the that the majority (at least have half) of the instances in each equivalence class should belong to the target set. Based on our previous work [6], $\alpha = 0.8$ can be a good value in the experiments in all methods. In all these methods, the swarm size is 30, the fully connected topology is used in PSO. $w = 0.7298$, $v_{max} = 6.0$, $c_1 = c_2 = 1.49618$ [14]. The maximum iteration is 200. In PSOPRSN, γ is set as 0.9, 0.8 and 0.7 to show the different importance of the classification performance and the number of attributes. Each method are conducted for 50 independent runs on each dataset.

To further examine the performance of PSOPRSE, two conventional filter dimension reduction methods (CfsF and CfsB) in Waikato Environment for Knowledge Analysis (Weka) [28] are used for comparison purposes. CfsF and CfsB are based on the correlation measure (Cfs) proposed by Hall [20], which measures the correlation between attributes and class labels. Cfs is implemented in Weka and it needs a search technique. Greedy search in Weka is selected as the search technique to perform both forward selection (CfsF) and backward selection (CfsB). The classification performance of the attributes obtained by CfsF and CfsB is calculated by DT.

V. Results and Discussions

Tables II shows the results of PSORS, PSORSN, and PSORSE. As the results of using 5NN are similar to that of DT and NB, they are not presented here due to the page limit. In the table, “PSORSN-0.9 , PSORSN-0.8 and PSORSN-0.7” show the results of PSORSN with γ values of 0.9, 0.8 and 0.7, respectively.

The classification performance of the selected attributes were evaluated by DT and NB on the test set of each dataset. In Table II, “All” means that all of the available attributes are used for classification. “Size” means the average number of attributes selected in the 50 independent runs. “Best” “Mean” and “StdDev” represent the best value, the average value and the standard deviation of the testing classification accuracies achieved by each algorithm in the 50 independent runs.

A. Results of PSORS

As shown in Table II, in almost all cases, PSORS reduced around one thirds of the available attributes in the datasets and obtained similar or even higher classification accuracies than using all attributes. In some cases, the classification performance of using all attributes is slightly better than the average classification performance of the selected attributes, but in most cases, the best classification performance is better

TABLE II
PSORS, PSORSN (WITH γ VALUES OF 0.9, 0.8 AND 0.7), AND PSORSE.

Dataset	Method	Size	DT		NB	
			Best	Mean±StdDev	Best	Mean±StdDev
Chess	All	36	98.5		87.89	
	PSORS	30.57	98.69	98.41±20.4E-2	91.17	88.67±1.61E0
	PSORSN-0.9	17.03	98.5	98.03±31.5E-2	94.55	92.25±1.22E0
	PSORSN-0.8	11.37	97.75	97.28±1.13E0	93.99	92.68±66.2E-2
	PSORSN-0.7	8.83	97.65	95.32±1.84E0	94.08	93.13±74.2E-2
	PSORSE	29.3	98.69	98.44±21.7E-2	91.46	88.61±1.6E0
Dermatology	All	34	82.79		95.9	
	PSORS	20.97	97.54	86.09±4.72E0	98.36	93.52±3.17E0
	PSORSN-0.9	8.83	96.72	74.86±7.89E0	95.08	80.93±5.85E0
	PSORSN-0.8	8.63	87.7	76.12±6.52E0	88.52	81.75±4.55E0
	PSORSN-0.7	7.83	95.08	76.89±6.93E0	90.16	80.44±5.88E0
	PSORSE	11.5	97.54	92.03±3.21E0	96.72	92.67±2.38E0
Lymph	All	18	75.51		87.76	
	PSORS	11.43	79.59	72.38±6.89E0	91.84	84.83±3.68E0
	PSORSN-0.9	5.17	71.43	65.78±5.69E0	83.67	78.16±1.68E0
	PSORSN-0.8	5.07	67.35	66.12±4.58E0	83.67	77.96±1.53E0
	PSORSN-0.7	5	67.35	67.35±31E-4	77.55	77.55±10E-4
	PSORSE	6.56	75.51	70.12±8.54E0	85.71	81.76±1.38E0
Soybeanlarge	All	35	81.94		90.31	
	PSORS	21.3	87.67	79.71±3.82E0	92.07	84.39±3.46E0
	PSORSN-0.9	10.37	80.18	72.36±3.91E0	81.94	76.8±3.47E0
	PSORSN-0.8	9.77	80.18	71.92±4.02E0	82.38	76.23±3.69E0
	PSORSN-0.7	9.43	79.74	72.28±4.1E0	85.9	76.04±4.47E0
	PSORSE	19.12	85.46	80.9±2.67E0	85.46	81.22±2.46E0
Waveform	All	40	74.79		79.71	
	PSORS	24.47	77.37	74.81±1.91E0	81.27	77.72±1.99E0
	PSORSN-0.9	8.23	76.71	68.8±2.87E0	75.75	69.86±3.52E0
	PSORSN-0.8	8	76.29	69.91±4.17E0	78.75	71.12±4.1E0
	PSORSN-0.7	7.97	73.29	68.38±4.65E0	75.03	69.59±4.51E0
	PSORSE	18.6	77.19	72.5±4.14E0	81.27	74.87±4.84E0
Statlog	All	36	86.39		82.61	
	PSORS	25.37	86.57	85.47±57.7E-2	82.61	82.06±27.9E-2
	PSORSN-0.9	13.8	86.57	84.9±73.3E-2	82.24	81.41±42.3E-2
	PSORSN-0.8	11.3	85.55	84.4±63.2E-2	82.24	80.45±1.23E0
	PSORSN-0.7	9.97	86.06	84.23±1.02E0	81.77	80.24±1.41E0
	PSORSE	20.04	86.81	85.42±76.8E-2	83.12	81.92±52.4E-2

than using all attributes. The results suggest that PSORS based on BPSO and probabilistic rough set theory can be successfully used to reduce the number of attributes needed for classification.

B. Results of PSORSN

According to Table II, by adding the number of attributes into the fitness function, PSORSN further reduced the number of attributes selected. PSORSN with a small γ selected a smaller number of attributes than with a relatively large γ. The reason is that a small γ in PSORSN means the number of attributes in the PSORSN is more important than a relatively large γ, and the classification performance is less important than with a large γ. Therefore, the fitness function (Fit_2) will lead PSORSN to search for the solution space with a smaller number of attributes. The results also show when the number of attributes decreases, the classification performance decreases in most cases. When PSORSN-0.8 or PSORSN-0.7, PSORSN could not achieve higher accuracies than using all attributes in most cases. This is consistent with our hypothesis in Section III-B. Without considering the size of the equivalence class, PSORSN could reduce the number of attributes in the reduct, but also reduce the generality of the reduct.

C. Results of PSORSE

According to Table II, in most cases, PSORSE selected less than half of the available attributes and obtained similar or higher even better accuracy than using all attributes. Although in some cases, the average classification performance of the selected attributes is slightly worse than using all attributes, their best classification accuracy is higher than using all attributes. The results suggest that PSORSE considering both the classification performance/representation power of the selected attributes and the number of equivalence classes can successfully reduce the number of attributes and maintain or achieve higher classification accuracy than using all attributes.

D. Comparisons Between PSORS and PSORSE

Comparing the results of PSORSE with those of PSORS, PSORSE achieved similar or higher classification accuracy than PSORS, but the number of attributes in PSORSE is always smaller or much smaller than in PSORS. For example, in the Dermatology dataset using DT as the classification algorithm, PSORS selected around 21 attributes from the 34 available attributes and its average classification accuracy is 86.09%. PSORSE further reduced around 50% of the attributes and improved the average classification performance to 92.03%. The main reason is that PSORSE considers the number of equivalence classes in the fitness function, which can further reduce/remove redundant or irrelevant attributes but keep the representation/classification power of the remaining attributes to achieve similar or higher classification accuracy than PSORS.

E. Comparisons Between PSORSN and PSORSE

In both PSORSN and PSORSE, the fitness functions consider both the classification power and the size of the attribute subset. By using different γ values, PSORSN usually obtained a smaller number of attributes, but for all the three learning algorithms (DT, NB and 5NN), PSORSE achieved higher or much higher classification accuracy than PSORSN, especially when PSORSN-0.7. The main reason is that PSORSN obtained a small number of attributes by directly considering the number of attributes, but without considering the size of the equivalence classes, the obtained attributes lost the generality and could not achieve good performance on unseen test data. Since the classification performance/representation power is usually considered more important than the number of attributes in dimension reduction problems, PSORSE can be regarded as a better dimension reduction approach than PSORSN.

Another advantage of PSORSE over PSORSN is that PSORSE does not need to predefine the parameter γ, which is typically difficult to determine. A larger γ means the classification performance is more important than a smaller γ, but the results in Table II show that the classification performance of PSORSN-0.8 is not always better than that of PSORSN-0.7, such as in the Lymph dataset. A possible reason is that PSORSN with PSORSN-0.7 further remove some redundant attributes, which reduce the complexity of the

TABLE III
RESULTS OF CFSF AND CFSB WITH DT AS THE LEARNING ALGORITHM

	Chess		Dermatology		Lymph	
	Size	Accuracy	Size	Accuracy	Size	Accuracy
CfsF	5	78.1	17	87.3	8	73.3
CfsB	5	78.1	17	87.3	8	73.3

	Soybeanlarge		Waveform		Statlog	
	Size	Accuracy	Size	Accuracy	Size	Accuracy
CfsF	12	80.5	32	72	5	71.62
CfsB	14	85.4	32	72	5	71.62

TABLE IV
CONTINUOUS DATASETS

Dataset	# Attributes	# Classes	# Instances
German	24	2	1000
World Breast Cancer -Diagnostic (WBCD)	30	2	569
Musk Version 1 (Musk1)	166	2	476
Semeion	256	2	1593
Madelon	500	2	4400

classification algorithms and slightly increase the classification performance. This suggests that the parameter γ, which is to balance the relative importance of the number of attributes and the classification performance, is problem-dependent and difficult to determine in advance.

F. Comparisons With Two Traditional Algorithms

Experiments using two traditional algorithms (CfsF and CfsB) for dimension reduction have been conducted using Weka. Experimental results are shown in Table III, where DT was used for classification. Comparing the experimental results of PSORS, PSORSN and PSORSE in Tables II with the results in Table III, we can observe that in almost all cases, PSORS, PSORSN and PSORSE achieved better or much higher classification accuracy than CfsF and CfsB, although CfsF and CfsB selected a smaller number of attributes.

Additionally, experimental results also show that using DT, NB and 5NN as the classification algorithms, the performance of PSORS, PSORSN or PSORSE show similar patterns. In most cases, the attributes selected by PSORS, PSORSN or PSORSE achieved similar or higher classification accuracy than using all attributes. This suggests that all the three filter methods based on PSO and probabilistic rough set are general to the three classification algorithms.

VI. FURTHER EXPERIMENTS ON CONTINUOUS DATASETS

Since all the discrete datasets we can find in UCI and other rough set related papers [9, 12, 11] have a small number of attributes, we use the data discretisation technique in Weka to pre-process the continuous data to discrete data. Five continuous datasets listed in Table IV were chosen from UCI and discretized. The five datasets were selected to have a large number of attributes (up to 500) and different numbers of classes and instances.

According to Table V, we can observe that all the three PSO and rough set based dimension reduction algorithms (PSORS, PSORSN and PSORSE) can be successfully used for the discretized continuous datasets, which have a large number of attributes. In all datasets, the number of attributes were significantly reduced. In most cases, by using the remaining small number of attributes, DT, NB and 5NN can achieve similar or higher classification accuracy than using all the original attributes.

TABLE V
PSORS, PSORSN (WITH γ VALUES OF 0.9, 0.8 AND 0.7), AND PSORSE.

Dataset	Method	Size	DT		NB	
			Best	Mean±StdDev	Best	Mean±StdDev
German	All	24	72.97		72.97	
	PSORS	16.9	74.17	71.79±1.29E0	79.28	76.49±1.94E0
	PSORSN-0.9	8.82	75.98	72.54±1.62E0	78.98	75.27±1.23E0
	PSORSN-0.8	8.08	75.98	72.49±1.13E0	78.38	75.39±1.09E0
	PSORSN-0.7	7.46	75.98	72.19±1.73E0	78.38	75.37±1.34E0
	PSORSE	13.24	75.68	71.72±1.65E0	78.98	75.11±1.31E0
WBCD	All	30	92.59		93.65	
	PSORS	18.74	96.83	93.82±1.38E0	97.88	95.72±1.36E0
	PSORSN-0.9	5.9	96.3	93.54±1.77E0	97.88	94±2.71E0
	PSORSN-0.8	5.22	96.83	94.02±1.56E0	98.94	94.22±2.68E0
	PSORSN-0.7	4.96	96.3	93.71±1.73E0	97.88	94.22±2.54E0
	PSORSE	8.98	96.3	93.04±1.59E0	96.83	92.85±1.56E0
Musk1	All	166	70.25		81.65	
	PSORS	100.32	80.38	71.63±3.72E0	80.38	75.47±1.92E0
	PSORSN-0.9	44.54	77.22	70.9±3.69E0	81.65	75.62±2.56E0
	PSORSN-0.8	44.54	77.22	70.9±3.69E0	81.65	75.62±2.56E0
	PSORSN-0.7	44.54	77.22	70.9±3.69E0	81.65	75.62±2.56E0
	PSORSE	80.98	79.11	70.9±3.83E0	79.11	75.67±1.92E0
Semeion	All	256	94.35		92.28	
	PSORS	158.9	94.35	92.61±80.6E-2	95.1	93.42±78.7E-2
	PSORSN-0.9	84.04	95.1	92.29±1.03E0	95.29	92.28±1.56E0
	PSORSN-0.8	84.04	95.1	92.29±1.03E0	95.29	92.28±1.56E0
	PSORSN-0.7	84.04	95.1	92.29±1.03E0	95.29	92.28±1.56E0
	PSORSE	143.08	94.35	92.43±91.7E-2	95.1	93.12±98.6E-2
Madelon	All	500	62.36		50.35	
	PSORS	299.78	83.37	75.18±7.26E0	61.89	57.44±2.16E0
	PSORSN-0.9	183.16	82.68	67.46±7.52E0	61.09	55.91±2.65E0
	PSORSN-0.8	183.16	82.68	67.46±7.52E0	61.09	55.91±2.65E0
	PSORSN-0.7	183.16	82.68	67.46±7.52E0	61.09	55.91±2.65E0
	PSORSE	299.78	83.37	75.18±7.26E0	61.89	57.44±2.16E0

Comparing the performance of PSORS, PSORSN with PSORSE, we can observe that their performance on the continuous datasets are generally similar to that on the discrete datasets. PSORS and PSORSE usually chose a larger number of attributes, but obtained higer classification accuracy than PSORSN. In most cases, PSORSE obtained a smaller number of attributes than PSORS and achieved similar or higher accuracy than PSORS. The attributes selected by the three algorithms are also general to the three classification algorithms (DT, KNN and NB) on the continuous datasets.

VII. CONCLUSIONS AND FUTURE WORK

This work aimed to propose a filter dimension reduction approach to classification problems with the expectation of reducing the number of attributes and maintaining or improving the classification performance over using all attributes. The goal was achieved by developing a new dimension reduction algorithm (PSOPRSE) based on PSO and probabilistic rough set. PSOPRSE aims to maximise the classification performance and minimise the number of attributes, where the classification performance is reflected by a probabilistic rough set theory measure and the number of attributes is reflected by the number of equivalence classes in probabilistic rough set theory. The performance of PSOPRSE was examined and compared with PSOPRS which maximises the classification performance only, PSOPRSN by adding the number of attributes in the fitness function, and two traditional filter dimension reduction algorithms. Experiments were conducted on six datasets of varying difficulty. Three classification algorithms, DT, NB and 5NN were used to test the generality of PSOPRSE. The results show that PSOPRSE outperformed PSOPRS in terms of both the number of attributes and the classification performance. Although the number of attributes in PSOPRSE is slightly larger than PSOPRSN and two traditional methods, PSOPRSE achieved better or much better classification performance than other methods mentioned above. Moreover, compared with PSOPRSN, PSOPRSE does not need to predefine a parameter to balance the relative importance of the number of attributes and the classification performance. Experimental results also show that PSOPRSE as a filter dimension reduction algorithm is general to the three different classification algorithms. As the discrete datasets include a small number of attributes, the performance of PSOPRSE, PSOPRS and PSOPRSN are further demonstrated on continuous datasets with a large number of attributes.

In the future, we will investigate ways to further reduce the number of attributes selected by PSOPRSE while maintaining its classification performance. We also intend to investigate multi-objective PSO and rough set theory for filter dimension reduction to search for the Pareto front of non-dominated solutions (attribute subsets) to provide more informative solutions to meet different requirements in real-world applications.

References

[1] I. Fodor, "A survey of dimension reduction techniques," Tech. Rep., 2002.

[2] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 4, pp. 131–156, 1997.

[3] S.-W. Lin and S.-C. Chen, "Psolda: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis," *Appled Soft Computing*, vol. 9, no. 3, pp. 1008–1015, 2009.

[4] K. Neshatian and M. Zhang, "Genetic programming for feature subset ranking in binary classification problems," in *European Conference on Genetic Programming*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 121–132.

[5] 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.

[6] L. Cervante, B. Xue, L. Shang, and M. Zhang, "A dimension reduction approach to classification based on particle swarm optimisation and rough set theory," in *25nd Australasian Joint Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 7691. Springer, 2012, pp. 313–325.

[7] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[8] Z. Pawlak, "Rough sets," *International Journal of Parallel Programming*, vol. 11, pp. 341–356, 1982.

[9] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459–471, 2007.

[10] Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 178, no. 17, pp. 3356 – 3373, 2008.

[11] H. Ming, "A rough set based hybrid method to feature selection," in *International Symposium on Knowledge Acquisition and Modeling (KAM '08)*, 2008, pp. 585–588.

[12] C. Bae, W.-C. Yeh, Y. Y. Chung, and S.-L. Liu, "Feature selection with intelligent dynamic swarm and rough set," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7026 – 7032, 2010.

[13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.

[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, 1997. Computational Cybernetics and Simulation.*, vol. 5, 1997, pp. 4104–4108.

[16] Y. Liu, G. Wang, H. Chen, and H. Dong, "An improved particle swarm optimization for feature selection," *Journal of Bionic Engineering*, vol. 8, no. 2, pp. 191–200, 2011.

[17] A. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 1100–1103, 1971.

[18] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *IEEE Transactions on Information Theory*, vol. 9, no. 1, pp. 11–17, 1963.

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

[20] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, 1999.

[21] H. Almuallim and T. G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artificial Intelligence*, vol. 69, pp. 279–305, 1994.

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

[23] K. Iswandy and A. Koenig, "Feature-level fusion by multi-objective binary particle swarm based unbiased feature selection for optimized sensor system design," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 365 –370.

[24] Y. Marinakis, M. Marinaki, and G. Dounias, "Particle swarm optimization for pap-smear diagnosis," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1645 – 1656, 2008.

[25] M. A. Esseghir, G. Goncalves, and Y. Slimani, "Adaptive particle swarm optimizer for feature selection," in *international conference on Intelligent data engineering and automated learning (IDEAL'10)*, 2010, pp. 226–233.

[26] H. N. A. Hamed, N. K. Kasabov, and S. M. Shamsuddin, "Quantum-inspired particle swarm optimization for feature selection and parameter optimization in evolving spiking neural networks for classification tasks," vol. 1, pp. 132–148, 2011.

[27] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

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